Biomedical Data Integration Framework for Glucose Level Prediction using Machine Learning Techniques

Haider Ali

20126384

School of Engineering, Computer and Mathematical Science

A thesis submitted to AUT University
in fulfilment of the requirements for the degree of
Doctor of Philosophy

2024

Supervisors
Dr. Samaneh Madanian, Dr. David White, Dr. Imran Khan Niazi

Abstract

Metabolic health conditions, characterized by elevated glucose levels, are significantly influenced by lifestyle factors such as diet, physical activity, and sleep. While Continuous Glucose Monitoring (CGM) devices provide essential insights into interstitial glucose (IG) levels, they often lack the broader physiological context necessary for a comprehensive metabolic assessment. This thesis investigates the potential of leveraging smartwatch data, specifically from devices like the Empatica E4, for interstitial glucose (IG) prediction through machine learning (ML). It makes four key contributions: (1) identifying taxonomies and methodologies for handling time-domain data in healthcare applications; (2) reviewing current digital biomarkers used for glucose prediction from smartwatch data and food logs; (3) comparing various ML models—including Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA), K-Nearest Neighbours (KNN), and Gaussian Naïve Bayes (GNB)—to determine the most effective algorithms for predicting IG levels from wrist-worn smartwatches and dietary logs; and (4) developing novel sleep-related features derived from smartwatch data that significantly enhance IG prediction accuracy.

A comprehensive systematic review of ML applications in time-domain electronic medical records identifies key models, features, and preprocessing techniques within the broader health data field. Additionally, a focused systematic review of digital biomarkers utilized in IG prediction highlights the necessity of comparing ML models and evaluating the utility of sleep biomarkers in glucose prediction. The investigation into most effective models for IG prediction reveals that that the RF model achieved the lowest Root Mean Squared Error (RMSE) of 9.04 mg/dL and an R-squared value of 0.84, whereas the GNB model exhibited the poorest performance, with an RMSE of 68.07 mg/dL. The ML models in this study use features measured from Empatica E4 smart watch data and food logs. The study then calculates sleep features derived from smart watch data, demonstrating that their inclusion in ML models, particularly DT, RF, and SVM, reduces the Mean Absolute Error (MAE) of predicted IG from 8.02 ± 0.22 mg/dL to 6.59 ± 0.33 mg/dL and enhances classification accuracy from 0.7988 ± 0.0183 to 0.8265 ± 0.0111, with statistical significance (p-value: 0.0001). These findings underscore the critical role of sleep metrics obtained from smart watch devices in improving the accuracy of glucose prediction models. By initially improving upon the literature with more effective use of smart watch sensor data, and subsequently incorporating novel sleep metrics derived from these devices, the study demonstrates a further refinement in predictive capabilities. Additionally, the application of explainable machine learning techniques, such as SHAP values, provides deeper insights into how various physiological factors

influence glucose levels (for example highlighting the effect of slow wave sleep and wake bouts on glucose prediction models), and partial dependence plots show how feature interactions are modelled by RF model. The broader implications of this work suggest that integrating multi-modal data from wearables into interpretable ML models could significantly advance the management of metabolic health conditions, offering more personalized and effective interventions.

Table of Content

List of Figures

List of Tables

Attestation of Authorship

I affirm that the content presented in this submission is entirely my original work. To the best of my knowledge and belief, it does not include any material previously published or authored by another person (except as explicitly acknowledged), nor has a significant portion of it been submitted for the attainment of any other degree or diploma from a university or other institution of higher learning.

……

Haider Ali (11/07/2024)

Copyright Statement

Scientific Contributions

Peer-reviewed journal publications and author contributions

Ali, H.; Niazi, I.K.; Russell, B.K.; Crofts, C.; Madanian, S.; White, D. Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning. Electronics 2023, 12, 554. https://doi.org/10.3390/electronics12030554.

Ali, H.; Niazi, I.K.; White, D.; Akhter, M.N.; Madanian, S. Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log. Electronics 2024, 13, 3192. https://doi.org/10.3390/electronics13163192.

Ali, H.; Niazi, I.K.; White, D.; Akhter, M.N.; Madanian, S. Digital Biomarkers from smart watches and food logs for Interstitial Glucose Prediction: A Systematic Review (Under Revew)

Ali, H.; Niazi, I.K.; White, D.; Akhter, M.N.; Madanian, S. Sleep Features for Prediction of Interstitial Glucose (Under Review)

Peer-reviewed conferences publications and author contributions

Ali, H, S. Madanian, N. Malik, D. White, B. K. Russel and I. K. Niazi, "Prediction of Interstitial Glucose Levels Through Smart watch Sensors Using Machine Learning," *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Nadi, Fiji, 2023, pp. 1-6, https://doi.org/10.1109/CSDE59766.2023.10487681

Ali, H.; Madanain, S.; White, D.; Akhter, M.N.; Niazi, I.K. From Smart watch Activity Trackers to Interstitial Glucose: Data to Insight—A Proposed Scientific Journey. *Proceedings of the 2024 Australasian Computer Science Week (ACSW '24*), Association for Computing Machinery, New York, NY, USA, 2024, pp. 61–64. https://doi.org/10.1145/3641142.3641154

Co Author Contributions

STUDENT AND SUPERVISOR APPROVALS					
By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.					
Student Name	Haider Ali	**Signature**		**Date**	11/7/2024
Supervisor Name	Samaneh Madanian	**Signature**		**Date**	11/7/2024

Chapter Number:	**2**
Manuscript Title:	Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning
Publication Status:	**Published**
Reference:	Ali, H.; Niazi, I.K.; Russell, B.K.; Crofts, C.; Madanian, S.; White, D. Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning. Electronics 2023, 12, 554. https://doi.org/10.3390/electronics12030554.

AUTHOR SURNAME:	CONTRIBUTION
Ali	Concept and design of the study, Data Analysis and interpretation, and Writing
Niazi	Concept and design of the study, Reviewing and editing
Russel	Concept and design of the study, Reviewing and editing
Crofts	Contribution of knowledge, Design of the study
Madanian	Concept and design of the study, Reviewing and editing
White	Concept and design of the study, Reviewing and editing

STUDENT AND SUPERVISOR APPROVALS					
By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.					
Student Name	Haider Ali	**Signature**		**Date**	11/7/2024
Supervisor Name	Samaneh Madanian	**Signature**		**Date**	11/7/2024

Chapter Number:	**3**
Manuscript Title:	Digital Biomarkers from smart watches and food logs for Interstitial Glucose Prediction: A Systematic Review
Publication Status:	**Submitted for Publication**

AUTHOR SURNAME:	CONTRIBUTION
Ali	Concept and design of the study, Data Analysis and interpretation, and Writing
Niazi	Concept and design of the study, Reviewing and editing
Madanian	Concept and design of the study, Reviewing and editing
White	Concept and design of the study, Reviewing and editing

STUDENT AND SUPERVISOR APPROVALS

By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.

Student Name	Haider Ali	**Signature**		**Date**	11/7/2024
Supervisor Nam	Samaneh Madanian	**Signature**		**Date**	11/7/2024

Chapter Number:	**4**
Manuscript Title:	Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log
Publication Status:	**Published**
Reference:	Ali, H.; Niazi, I.K.; White, D.; Akhter, M.N.; Madanian, S. Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log. Electronics 2024, 13, 3192. https://doi.org/10.3390/electronics13163192.

AUTHOR SURNAME:	CONTRIBUTION
Ali	Concept and design of the study, Data Analysis and interpretation, and Writing
Niazi	Concept and design of the study, Reviewing and editing
White	Concept and design of the study, Reviewing and editing
Akhter	Concept and design of the study, Reviewing and editing
Madanian	Concept and design of the study, Reviewing and editing

STUDENT AND SUPERVISOR APPROVALS

By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.

Student Nam	Haider Ali	**Signature**		**Date**	11/7/2024
Supervisor Name	Samaneh Madanian	**Signature**		**Date**	11/7/2024

Chapter Number:	**5**
Manuscript Title:	Utility of sleep features for interstitial glucose prediction
Publication Status:	**Submitted for Publication**

AUTHOR SURNAME:	CONTRIBUTION
Ali	Concept and design of the study, Data Analysis and interpretation, and Writing
Niazi	Concept and design of the study, Reviewing and editing
White	Concept and design of the study, Reviewing and editing
Akhter	Concept and design of the study, Reviewing and editing
Madanian	Concept and design of the study, Reviewing and editing

Acknowledgements

The authors would like to acknowledge the kind support provided by ECMS fee scholarship for the duration of the PhD. The kind support provided by the New Zealand College of Chiropractic in the form of a scholarship. The author would like to acknowledge the kind and meaningful insights provided by Dr. Samaneh, Dr David and Dr Imran. The author would like to extend his gratitude to Malik Naveed Akhter, Usman Sheikh, Usman Ghani and Alamdar Hussain for the spirited discussions and the support for the project. I would also like to extend my gratitude to Samana, Maria, Sehrish, Asad and Masooma for their endless contributions to my learning journey. I would also like to thank my mother who taught herself English to teach me, and my father who taught me the value of honest work.

List of Abbreviations

Abbreviation	Full Form
AHA	American Heart Association
ACC	Accelerometer
ADABOOST	Adaptive Boosting
ANS	Autonomic Nervous System
BVP	Blood Volume Pulse
CDC	Centre for Disease Control
CGM	Continuous Glucose Monitoring
DACIA	Data, Aggregation, Contextualization, Integration and Action
DT	Decision Tree
ECG	Electrocardiogram
EDA	Electrodermal Activity
EMR	Electronic Medical Record
HR	Heart Rate
HRV	Heart Rate Variability
IBI	Inter Beat Interval
IG	Interstitial Glucose
KNN	K-Nearest Neighbour
LassoCV	Lasso Cross Validation
ML	Machine Learning
NREM	Non- Rapid Eye Movement
PPG	Photoplethysmography
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF	Random Forest
REM	Rapid Eye Movement
SHAP	Shapely Additive Explanation
SLR	Systematic Literature Review
SOT	Sleep Onset Time

TIR	Time in Range (IG)
TST	Total Sleep Time
WASO	Wake After Sleep Onset
WT	Wake Time
XGBoost	Extreme Gradient Boosting
SVM	Support Vector Machine
PDP	Partial Dependence Plot
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
CT	Computerized Tomography
TM	Translational Medicine
NLP	Natural Language Processing
LASSO	Least Absolute Shrinkage and Selection Operator
TIR	Time in Range
MEMS	Micro-Electro-Mechanical System
GSR	Galvanic Skin Response
SCR	Skin Conductance Response
CLAID	Closing the Loop on AI & Data Collection
MyPHD	My Personal Health Dashboard
WT	Wake Time
WASO	Wake After Sleep Onset Time
TST	Total Sleep Time
SOT	Sleep Onset Time
Wake	W
N1, N2 and N3	Non-Rapid Eye Movement 1,2 and 3
SWS	Slow Wave Sleep
REM	Rapid Eye Movement
PSG	Polysomnography
t-SNE	t-Distributed Stochastic Neighbour Embedding
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error

MSLE	Mean Squared Logarithmic Error

1 Introduction

1.1 Context

Metabolic disorders affect large populations worldwide, with more than 537 million adults globally diagnosed with diabetes as of 2021, and this number is projected to rise to 783 million by 2045 (Gupta et al., 2024). Metabolic disorders are primarily characterized by elevated glucose levels and are influenced by a range of factors, including genetics, lifestyle, and pre-existing conditions like obesity or hypertension (Umpierrez et al., 2024). Among the risk factors that influence metabolic health, lifestyle is the only one that can be directly modified. Key lifestyle factors influencing metabolic health include diet, sleep, and physical activity. For instance, poor sleep quality and insufficient physical activity have been associated with impaired glucose metabolism and increased risk of metabolic disorders (Crofts et al., 2016; Spiegel et al., 1999; St-Onge et al., 2016).

It is well-documented that individuals with pre-existing conditions face a higher risk of developing metabolic disorders when exposed to unhealthy lifestyle such as poor diet and lack of physical activity (F. B. Hu et al., 2001). However, research shows that modest lifestyle changes, for example, change of 5-10% of body weight or engaging in 150 minutes of physical activity per week can substantially reduce the risk of developing type 2 diabetes in individuals with prediabetes (Knowler et al., 2002; Kurnik Mesarič et al., 2023). Furthermore, lifestyle interventions that address diet, physical activity, and sleep have been shown to improve glucose control and reduce the need for medication in patients with type 2 diabetes (Golovaty et al., 2023; Patel et al., 2006; Tuomilehto et al., 2001).

Metabolic disorders are typically classified into two categories: Metabolic Syndrome and Metabolic Disease. Metabolic Syndrome refers to a cluster of reversible risk factors that, if left untreated, can progress into chronic metabolic diseases like cardiovascular disease, stroke, and type 2 diabetes (Mohamed et al., 2023). According to the American Heart Association (AHA), metabolic syndrome affects approximately 23% of adults worldwide (Alberti et al., 2009). For example, prediabetes—a metabolic syndrome where fasting blood glucose levels are between 100 and 125 mg/dL—affects 96 million U.S. adults, representing over 37% of the population (CDC, 2024b). If left untreated, prediabetes can develop into type 2 diabetes, defined by fasting glucose levels exceeding 125 mg/dL. Each year, approximately 5-10% of individuals with prediabetes progress to type 2 diabetes (J. Hu et al., 2023; Tabák et al., 2012b).

In New Zealand, approximately 268,000 people are diagnosed with type 2 diabetes, and the prevalence is projected to rise, especially in high-risk groups such as Māori and Pacific populations (Daly et al., 2024). The New Zealand Ministry of Health highlights that metabolic syndrome is a growing concern, significantly contributing to the country's high rates of cardiovascular disease and type 2 diabetes (Mustafa et al., 2024). Additionally, prediabetes affects an estimated 25% of New Zealand adults, representing approximately 1.1 million people.

The high prevalence of prediabetes often goes undetected. For example, in the United States of America, approximately 80% of individuals with prediabetes are unaware of their condition, according to Centre of Disease Control (Dugani et al., 2024) . This lack of awareness is concerning because prediabetes can be reversed through early lifestyle interventions, such as dietary changes, increased physical activity, and weight management. Research shows that these modifications can significantly reduce the risk of progression to type 2 diabetes (Amelia et al., 2024; Jumpertz Von Schwartzenberg et al., 2024; Länsitie et al., 2021). As prediabetes and diabetes rates continue to rise globally, including in New Zealand and the U.S.—where diabetes now affects over 37.3 million people as of 2022—proactive management is crucial to reversing these trends (Eades et al., 2024). A significant component of such proactive management is awareness of levels of glucose in the body (Jain, 2024)

For both diabetes and prediabetes, elevated glucose levels, known as hyperglycaemia, pose significant health risks if left unmanaged. Chronic hyperglycaemia, a common complication in diabetes, can lead to potential long-term damage to critical organs such as the heart, kidneys, nerves, eyes, and blood vessels (Umpierrez et al., 2024). Hyperglycaemia is reported to increase the risk of cardiovascular disease by two to four times (Mouri & Badireddy, 2023).

While hyperglycaemia is a common result of metabolic system impairment, hypoglycaemia, defined as a blood glucose level below 70 mg/dL, can also have serious consequences. Symptoms of hypoglycaemia include shakiness, sweating, dizziness, confusion, and, in severe cases, seizures or unconsciousness (McCall et al., 2023). Repeated episodes of hypoglycaemia can result in hypoglycaemia unawareness, a condition in which individuals fail to recognize the warning signs of low blood sugar, leading to potentially life-threatening situations if not treated immediately (Heller et al., 2019).

Another condition that affects people with metabolic conditions is insulin resistance. Insulin is a hormone that regulates blood glucose levels by facilitating the uptake of glucose into cells, while insulin resistance occurs when cells become less responsive to insulin, leading to elevated blood sugar levels (X. Zhao et al., 2023). Insulin resistance can lead to serious health conditions such as type 2 diabetes, cardiovascular disease, metabolic syndrome, non-alcoholic fatty liver disease, polycystic ovary syndrome, chronic inflammation, certain cancers, and cognitive decline (Roden et al., 2024).

All these complications related to metabolic conditions are tied to glucose levels in the body. Hence, it is crucial to monitor glucose levels in the body to manage and prevent these complications. One classical method of measuring glucose is the fasting glucose test, where blood is sampled after a period of fasting. This test is used to assess long-term glucose control and is called Haemoglobin A1C (HbA1C). HbA1C measures the average blood glucose levels over the past two to three months by calculating the percentage of haemoglobin (a protein in red blood cells) that is coated with sugar (glycated) (Little & Sacks, 2009). This test offers a clearer picture of how well blood sugar has been managed over time, making it a key tool for diagnosing and monitoring diabetes and prediabetes (Sherwani et al., 2016). For instance, an HbA1C level below 5.7% is considered normal, while levels between 5.7% and 6.4% (100-125 mg/dL) indicate prediabetes and a level of 6.5% (>125 mg/dL) or higher is diagnostic of diabetes (Nicolaisen et al., 2023) .

While this tracking of HbA1C is crucial for monitoring metabolic health, it has some shortcomings. It does not evaluate the instantaneous state of metabolic health, as it measures the long-term control of diabetes. It also requires frequent sampling and handling of blood to track the trends of glucose changes during the day (Poolsup et al., 2013). To overcome these shortcomings, Continuous glucose monitoring (CGM) systems have come to the fore. CGMs generally track the glucose level in the interstitial fluid (Jackson et al., 2021). They comprise a small thread that penetrates interstitial fluid, measuring the interstitial glucose (IG) levels every one to five minutes (Huang et al., 2023). IG values in CGMs are stored on the device for up to 8 hours. Frequent IG monitoring using CGMs helps manage metabolic disorders by providing real-time insights into glucose levels frequently throughout the day and night (Bialasiewicz et al., 2009).

CGMs facilitate the estimation of various glycaemic control markers. One such marker is Time in Range (TIR), reflecting the percentage of time a person’s glucose levels remain within a target range, typically between 70 and 180 mg/dL (Beck et al., 2019). Maintaining glucose within this range is critical for reducing the risk of complications like

cardiovascular disease, neuropathy, and retinopathy, which are exacerbated by hyperglycaemia (Battelino et al., 2019). Another marker of glycaemic control measured with CGM data is Glycaemic Variability (GV), measuring fluctuations in glucose levels. High GV is linked to oxidative stress and inflammation, both of which contribute to long-term organ damage, such as cardiovascular disease (Monnier et al., 2008). CGMs can also track postprandial spikes (glucose spikes after meals), which are linked to increased risks of cardiovascular events (Kroeger et al., 2021). Managing postprandial spikes can help prevent the progression of complications, especially in individuals with prediabetes or diabetes (Ceriello, 2005). In addition, CGMs detect low glucose events (hypoglycaemia), providing timely alerts when glucose levels fall below 70 mg/dL, thus preventing severe outcomes such as seizures or unconsciousness (Yun & Ko, 2015). On the other end, CGMs also monitor hyperglycaemia (Yeung et al., 2024). Glucose trend arrows can also be measured with CGM data that help determine whether glucose levels are rising or falling, allowing for proactive adjustments in diet, physical activity, or medication to prevent dangerous excursions (Ahmann et al. 2018).

While the glucose control markers measured by CGMs are very important for people with metabolic conditions, CGMs are not routinely used by healthy individuals. However, tracking glucose levels for the healthy can provide valuable insights into how the body responds to different foods, stress, and exercise (Dehghani Zahedani et al., 2021). For example, studies have shown that even in non-diabetic individuals, postprandial glucose spikes above 140 mg/dL can occur after consuming high-carbohydrate meals, which, over time, can impair insulin sensitivity (Jackson et al., 2021). For athletes or those interested in performance, real-time glucose data is particularly important. Studies have demonstrated that real-time glucose monitoring helps athletes fine-tune their nutrition and recovery strategies by understanding how their glucose levels fluctuate during different types of exercise and recovery periods (Jeukendrup, 2014).

Even the highly frequent markers provided by CGMs are insufficient on their own to fully assess the impact of different lifestyle changes on glycaemic control. These lifestyle changes can be measured using smart watches, which have become more prevalent in recent years. With an estimated 200 million users worldwide as of 2022 and projections of global shipments reaching 253 million by 2025, smartwatches are being increasingly integrated into everyday life (Masoumian Hosseini et al. 2023). The widespread use of smartwatches, combined with advancements in sensor technology, offers unprecedented potential for metabolic health management. Modern smartwatches now monitor heart rate variability (HRV), blood oxygen levels (SpO2), electrodermal activity (EDA), sleep patterns, and even provide on-demand ECG and blood pressure tracking

(Ha et al., 2023). These features allow for continuous tracking of key physiological markers that are closely linked to metabolic conditions.

This thesis is aimed at predicting IG levels using smart watch data. This work predicts IG from Empatica E4 smart watch. To model the values of IG from an E4 device requires the knowledge of the exact relationship between the smart watch data and the IG values. Recent advances in Machine Learning (ML) techniques have been used to learn this relationship from the examples of inputs and outputs (labels) (Dehghani Zahedani et al., 2021). It typically involves the conversion of data into features. Features are the data attributes that are more closely related to the labels. This thesis compares the efficacy of different ML models in finding the relationship between E4 data and the IG values. This work also adds to the body of work on new sleep-related features measured using smartwatch data that help improve the performance of ML models in predicting IG values. This work also explains the outcomes of the ML models based on the relative feature importance of the input features using shapely additive explanations (SHAP) values.

1.2 Motivation

The motivation behind this research stems from the growing global prevalence of metabolic disorders, particularly diabetes, which currently affects over 537 million adults and is projected to rise to 783 million by 2045 (Gupta et al., 2024). These disorders are primarily influenced by modifiable lifestyle factors such as diet, sleep, and physical activity, and addressing these through early interventions has been shown to reduce the progression of prediabetes to type 2 diabetes significantly (Knowler et al., 2002; Kurnik Mesarič et al., 2023). However, to assess the effectiveness of these interventions at an individual scale, there is a critical need for continuous and non-invasive monitoring tools that provide real-time insights into glucose regulation. Traditional glucose monitoring methods, such as HbA1C and fasting glucose tests, have limitations in capturing the day-to-day variability in glucose levels, which is essential for preventing the complications associated with hyperglycaemia and hypoglycaemia (Poolsup et al., 2013; Sherwani et al., 2016).

The emergence of CGM systems makes glucose helps the management of metabolic conditions by providing real-time IG levels. However, these systems are also invasive and expensive, limiting widespread adoption. On the other hand, the growing use of data-rich smartwatches, with over 200 million users worldwide, presents an opportunity for utilisation in IG level tracking (Masoumian Hosseini et al., 2023). This research is

motivated by the potential to leverage the physiological data collected by smartwatches, such as HRV, skin temperature, accelerometers, and electrodermal activity (EDA), to predict IG levels using ML models. This non-invasive, accessible approach has the potential to empower individuals, particularly those at high risk of developing diabetes, by offering continuous monitoring and personalized feedback that encourages lifestyle modifications, improving glucose control, and reducing the risk of complications.

The research further aims to enhance the precision of glucose predictions by developing and comparing different ML models, utilizing features measured by the Empatica E4 smartwatch. This exploration is not only significant for advancing the accessibility of glucose monitoring but also holds broader implications for public health, particularly in populations disproportionately affected by metabolic disorders, such as the Māori and Pacific communities in New Zealand.

This research also measures sleep related features from smart watch data and uses it in explained models to predict IG levels, hence allowing for assessing the effect of sleep on IG regulation at an individual scale, thus enabling targeted interventions related to sleep.

1.3 Contributions

This thesis has four main contributions; identifying taxonomies and research related to time domain data in healthcare applications, identifying current state of the art for digital biomarkers used for the prediction of glucose values using smart watch data and food logs, comparison of different ML models in predicting the IG values from wrist worn smart watch sensors and food logs, developing novel sleep features used in predicting IG values from smart watches and food logs.

This thesis has the following main contributions:

1. The first contribution is a literature review titled: Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning is broad in nature and is not limited to the problem of IG prediction from smart watch sensors. This review adds to the body of literature the different methods used for handling time domain healthcare data, the research questions that can be answered using time domain healthcare data and methods used to address these questions. The rationale behind this work is first to consolidate the disparate body of work for future researchers to understand various components of healthcare solutions based on time domain data. This research is focused on finding these

solutions based on time domain data, as most sensors in healthcare store time logged values. Sensors measure physiological processes with the help of transducers, for example, a Photoplethysmography (PPG) sensor usually contains a light emitting diode (LED) that shines the light through the skin and a photo detector that measures the amount of reflected light, this amount of reflected light is proportional to the amount of blood, hence PPG sensor value in time is used to measure blood volume pulse (BVP), heart rate (HR), oxygen saturation in blood, the rate of change of these values are used to measure different markers of sympathetic nervous system (SNS) such as HRV and IBI. The values measured are not without errors and noise and are an indirect measure of the physiological processes they measure. The number of samples a sensor measures in a second is termed its sampling frequency. This noise in the time series data calls for a series of steps from sampling the data, storing the data, filtering out the noise of the data and applying some mathematical operations on the data to find the physiological events of interest. Since this data is in the healthcare domain, there are certain operations that need to be performed on the data to ensure compliance with legal and privacy principles. ML methods are sometimes used to learn relationships between sensor values and physiological events of interest. Other research questions can also be asked from time domain healthcare data such as increase or decrease in health. This review thus identifies different axes or taxonomies alongside the application of time-domain healthcare data from electronic medical records (EMRs). This review informs various aspects of handling data to answer research questions, and the methods used in related works to answer them.

2. The second contribution in this work is a more targeted literature review titled: Digital Biomarkers from Smart Watches and Food Logs for Interstitial Glucose Prediction: A Systematic Review. It is focused on identifying different digital biomarkers that are used in predicting IG values from smart watches and food log data. Digital biomarkers are defined as: objective, quantifiable, physiological measures that can be measured using digital means or sensors (Anmella et al., 2024). These digital biomarkers are contrapositives to the features used in ML models for healthcare applications. Features are the aspects of input data that help the ML models learn the relationship between input and output labels. Consequently, digital biomarkers can serve as features to learn the relationship between input data and output label (Data àDigital Biomarkersà Model). A recent work (Daniore et al., 2024) proposed a framework containing guiding questions for Data collection, Aggregation, Contextualization, Interpretation and Action (DACIA) processes to develop digital biomarkers. Chapter 3 of this thesis

reviewed past research that used ML models, smart watch and food log data. The biomarkers identified in the literature searched were classified into the autonomic nervous system (ANS), food, activity, and statistical and circadian features informed by the patterns found in the literature. For each of the classes of features, the guiding questions of the DACIA framework are answered. This answer informs the data handling, preprocessing, filtering, windowing and feature calculation processes. This work then identifies different processes within data to biomarker pipelines and reports them. The rationale behind Chapter 3 is to list and analyse different operations performed on smart watch data to develop robust pipelines. The answer to DACIA questions allows the user to develop novel methods within each stage of data processing and compare them to the state of the art.

3. Another component of this thesis is paper titled: Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log. It is a comparison of different ML models in glucose prediction using biomarkers measured from smart watch sensors (Chapter 4). Most of the biomarkers identified in Chapter 3, can be used by different ML models in IG prediction tasks. Each model has unique advantages and limitations based on the statistical biases induced in the models. By systematically comparing ML models using a combination of smart watch sensor data and food logs, this work aims to identify the most effective algorithms for predicting IG levels from smart watches. The best models for IG prediction from food logs and smart watches are identified, their hyperparameters optimized and the models explained using SHAP values. Explainable ML models that combine different features measured from data can offer crucial insights into how lifestyle choices impact glucose levels, helping patients and clinicians make early interventions. For healthy individuals, these models can be used for preventive care, providing feedback on how different physiological and activity patterns influence metabolic health and can be used for suggesting adjustments to prevent the development of metabolic syndrome. In sum, comparing ML models ensures that the most appropriate predictive system is selected for each patient profile, enhancing early detection, personalized care, and long-term management of metabolic conditions.
4. The next component of this PhD work is another paper titled: Utility of Sleep Features in ML for Prediction of Interstitial Glucose. It is a design of novel sleep features from HR and accelerometers and using them to predict IG values (Chapter 5). Chapter 5 identifies relevant sleep features that can be measured from HR and accelerometer sensors. It then lists all the open-source methods used to measure identified sleep parameters. These sleep parameters include

total sleep time (TST), wake after sleep onset (WASO), sleep onset time (SOT), wake time (WT), and sleep stages (rapid eye movements (REM), non-rapid eye movements (NREM (N1, N2, N3)) and Wake). Sleep stages and sleep quality have been demonstrated to influence glucose regulation and metabolic health (Byun et al., 2020; Spiegel et al., 1999) but no work, to the authors' knowledge, has used HR and accelerometers to calculate these features and use them to predict IG values. This work modifies the features to define more novel features descriptive of IG changes. Sleep features are then used to train different ML models. The model performance has increased with the addition of sleep features demonstrating their effectiveness. Like Chapter 4, in Chapter 5 as well the best performing ML models are explained using SHAP values. The SHAP values for different features demonstrate which features are most influential in predictions made by the models.

The clinical application of this PhD work lies in the comparison of ML models and proving that tree-based models outperform other ML models in the task of predicting IG values in Chapter 4. This finding can help future researchers identify the baseline models against which they can develop solutions. The performance of tree-based models is used to argue about the underlying distribution of the data as well. The outperformance of tree-based models is based on statistical tests such as Nemenyi post hoc analysis to identify the statistically significant changes.

he development of sleep-related features from HR and accelerometer data, which significantly improved the accuracy of glucose prediction models for individuals at various stages of metabolic health (Chapter 5). Sleep plays a critical role in metabolic regulation, and sleep disruptions—whether due to poor sleep quality, irregular sleep patterns, or insufficient sleep—have been shown to impair insulin sensitivity and contribute to glucose dysregulation. By incorporating features such as HRV, sleep duration, and sleep fragmentation derived from smart watch sensors, this research introduces a novel and important layer of predictive accuracy for ML models for IG prediction.

For pre-diabetic individuals, where metabolic dysfunctions may not yet be severe but pose a significant risk, this approach allows for early detection of glucose irregularities triggered by poor sleep. This helps in informing the lifestyle changes that can help lower glucose levels or plan interventions targeted at different aspects of sleep quality.

1.4 Significance of the Research

Integrating smart watch technology with ML models to predict IG values has important potential clinical applications for metabolic disorders. By utilizing the Empatica E4 smartwatch to continuously monitor physiological markers such as HRV, T, EDA, and physical activity, this thesis proposes a non-invasive and user-friendly approach to glucose monitoring. Traditional CGM systems, while effective, can be cumbersome and costly, limiting their accessibility and widespread adoption. In addition to the predictive ability of the models, these models can be used in conjunction with CGM and smart watches to explain the changes in IG values. Using widely available smartwatches for glucose level monitoring makes it more accessible to a larger population, which can potentially help people at risk of diabetes track their levels of glucose. The SHAP explanations can help identify the source of high glucose excursions potentially encouraging lifestyle changes.

In clinical practice, utilizing ML models with smart watch devices to predict IG levels can potentially improve the accuracy and responsiveness of diabetes management (Zahedani et al., 2023). While this study develops models for IG prediction on a limited dataset, it demonstrates the development of explained ML models for IG prediction. With access to how the input features are affecting the output values, healthcare providers can suggest personalized advice, enabling timely interventions to prevent acute complications such as hyperglycaemia and hypoglycaemia. This level of detailed monitoring and prediction supports personalized medicine, where interventions are tailored to the individual's unique physiological responses and lifestyle factors. Additionally, classifying IG values into personalized categories facilitates better patient education and self-management, empowering individuals to take proactive steps in maintaining their metabolic health.

For developers of healthcare solutions, the comparative analysis of different ML models in IG prediction from smart watches helps identify the most effective methods. The statistical comparison of these ML models using performance metrics ensures the reliability and accuracy of the predictions, which is crucial for clinical trust and adoption. Using technology to develop solutions for metabolic health is crucial due to the widespread prevalence of prediabetes and diabetes, as early detection and intervention can significantly prevent the progression to more severe metabolic disorders.

From a public health perspective, the scalability and cost-effectiveness of using smartwatches for glucose monitoring can have a significant impact on managing the growing burden of metabolic disorders globally and within specific populations such as those in New Zealand. High-risk groups, including Māori and Pacific populations, who are disproportionately affected by type 2 diabetes, can particularly benefit from accessible and continuous monitoring solutions. By enabling large-scale deployment of smartwatch-based CGM, healthcare systems can implement more effective screening and prevention programs, ultimately reducing the incidence of diabetes-related complications and associated healthcare costs.

Moreover, the insights gained from this research contribute to the broader understanding of the relationship between lifestyle factors and glucose metabolism. By identifying the most influential physiological markers through ML models, this study provides a foundation for developing integrated health platforms that combine CGMs with other health metrics tracked by smartwatches. Such integration fosters a holistic approach to metabolic health, where multiple aspects of an individual's lifestyle and physiology are considered in concert to optimize health outcomes.

1.5 Research Objective

The primary objective of this research is to develop and evaluate ML models for predicting IG levels using data collected from smartwatches and food logs. Specifically, the study aims to assess the effectiveness of various ML algorithms in accurately predicting IG values, while also investigating the impact of including sleep-related features derived from smartwatch data on model performance. By incorporating SHAP analysis, this research further seeks to enhance the interpretability of the models, providing insights into the most significant predictors of glucose levels. Ultimately, the goal is to offer a non-invasive, accessible, and personalized solution for monitoring glucose.

1.6 Statement of Purpose

This research aims to understand the feasibility of the use of ML models using smart watch sensors in predicting IG levels. The work centres around four key research questions for their potential to augment the current body of knowledge on glucose prediction. In the subsequent sections, this work explores the possibilities of ML as a mean for IG prediction through the following research questions:

1. What is the state of the art of data processing and ML applications in time domain healthcare data?

2. What is the state of the art in digital biomarker design in predicting interstitial glucose from smart watch sensors?
3. How do different ML models compare in predicting IG from smart watch and food log data?
4. How does sleep parameters measured using HR and accelerometer data affect the performance of different ML models in predicting IG levels?

The first research question aims to understand the different methods currently used in storage, processing, filtering and calculating different aspects of time domain healthcare data. The answer to this question adds to the body of work by consolidating different processes, including the privacy and legal standards that need to be complied with in developing healthcare solutions from time domain data.

The second research question delves into the exploration of different digital biomarkers and how they are used in predicting IG values. The answer to this question will inform the different operations performed on the data in accordance with the DACIA framework. This helps researchers in developing new digital biomarkers by following the same processes. It will also help researchers develop new digital biomarkers by using novel data collection, aggregation, contextualization, integration, and action methods. This also lists the preprocessing steps used by different researchers; this can help in developing standardized methods or rules for preprocessing in the context of digital biomarkers.

The third research question shifts from examining the state of the art and applying different ML models in predicting glucose levels from smart watch sensors. This work compares the efficacy of different ML models, including Bayesian models, linear models, non-parametric models (KNN), tree models (DT, RF, XGBOOST and maximal margin models (SVM). The answer to this question can inform the underlying distributions of the data, the nature of the errors and the outliers in the data. The outperformance of the tree model shows the presence of influential outliers in the data. The best performing models are trained and their hyperparameters are tuned and reported using Bayesian Optimization. This will help future work in designing efficient baseline models. The models are compared based on statistical tests in accordance with (Rainio et al., 2024).

The final research questions assess the performance of sleep features in the prediction of IG values from smart watch data. The answer to this question is beneficial in two aspects: it increases the performance of the models by considering the effect sleep has on the IG values. Secondly, a SHAP explanation of the model can help participants understand the effect sleep had on their glucose levels and can adjust the sleep aspects of their lifestyles. The relative importance of the features shows that sleep features are

almost as important as food features in the ML models, signifying the importance of sleep models. In addition to that these sleep features are developed based on open-source methods. The methods are reported, and the performance of the models by the addition of sleep features is reported. Statistical tests are carried out to make sure that the increase in model performances is statistically significant.

1.7 Thesis methodology

The first research question is answered by conducting a systematic literature review (SLR). The SLR is carried out by first identifying the relevant records from the scientific databases including PubMed, Scopus and Web of Science using a search term. The searched articles are then filtered using preferred reporting items for systematic review and meta-analysis (PRSIMA) using an exclusion and inclusion criteria. The studies identified are then reported with the help of a synthesis table. The identified taxonomies are clearly delineated and explained.

The second research question is also answered based on another SLR. The information from SLR in Chapter 3 is gathered by identifying records and filtering using a PRISMA process. The information is synthesized using the DACIA guiding questions. The resulting answers help find methods and processes used in digital biomarkers designed for predicting glucose levels from smart watch sensors.

The answer to the third research question is found using the following methodology. A public dataset (D1) containing glucose labels from Dexcom CGM and Empatica E4 smart watch are used (Cho et al., 2023a). This data is first pre-processed, and then converted into features for prediction windows. Different ML models are trained based on 70% of available data and tested based on 30% of remaining data. The results of the models are compared based on statistical tests according to (Rainio et al., 2024). The predictions of the best performing ML models are explained using explainable AI techniques including SHAP values and partial dependence plots (PDP). This helps inform which models perform better than others and how these models use input features in predicting IG values.

For answering the final research question, an SLR is first conducted to identify sleep features that are estimated using accelerometers and HR features. After identifying the features, accelerometer and HR data is pre-processed. These features include sleep stages (NREM (N1, N2, N3), REM and W) and TST, SOT, WP and WASO. For sleep stage classification, a public dataset (D2) having apple watch data containing HR, ACC and PSG labels is used. A random forest (RF) model is trained on this dataset to predict sleep stages from the HR and ACC data. To do inference on this model, data columns

from D1 are modified to make them compatible with D2 and 30 second windows corresponding to sleep epochs are created. After which sleep stages are estimated. These sleep stages alongside other sleep features are then tested for correlation with the IG values. To improve the correlation between IG and estimated sleep features, sleep features are transformed using statistical transformations. These transformations include log, box-cox, yeo johnson and rank based inverse transforms. The results of these transformations are also then correlated with the IG values and transformations with statistically significant correlations are selected. Sleep features are also transformed by finding the percentage of each sleep stage and difference from recommended sleep stage. These features are then input into a series of ML models including adaptive boosting (adaboost), decision tree (DT), KNN, Lasso cross validation (LassoCV), RF, SVM and XGBoost. The models are first provided with features without the sleep features, then all sleep features are added including all the modifications, then only transformed features are added. The performance increases are verified with paired t-tests and the models are compared based on Friedman's test and Nemenyi post-hoc analysis. The results indicate that sleep features significantly increase the prediction performance of ML models on unseen data. The best performing ML models are explained for all features as well as sleep features using SHAP values allowing for examination of which features are most crucial in predicting IG values from smart watch data. The results of the SHAP analysis highlight the importance of sleep features especially WASO and N2 in ML based IG prediction from smart watches.

1.8 Thesis organisation

This thesis unfolds as a systematic progression of studies, organized into Chapters that build upon each other, as illustrated in Figure 1.1. Chapters 2-5 have been prepared for publication in distinct quality peer-reviewed journals and may exhibit some recurring content. Within this thesis each Chapter will be introduced with prefaces, retaining its original published form.

Chapter 2 conducts a comprehensive review of existing scientific literature on Review of Time Domain EMR Taxonomies in application of ML. It establishes a context for the design requirements of time domain data handling and feature calculation, emphasizing the need for a distinct approach for time domain healthcare data. Taxonomies within this broad field are highlighted and the processing steps to answer research question based on time domain healthcare data are listed and compared analytically.

In Chapter 3, Digital Biomarkers from smart watches and food logs for Interstitial Glucose Prediction: A Systematic Review different digital biomarkers that are used in literature for predicting IG values from smart watch sensors are listed. This review informs various

preprocessing steps, modelling steps and importance of digital biomarkers in prediction of IG values.

Chapter 4 delves into the comparison of different ML models in predicting IG values from smart watches. A public dataset is used. The data is then filtered using the required preprocessing steps and the windowed for 5-minute prediction windows. ML models are also used to predict IG values from the input features. The performance of the models is compared based on the relevant performance metrics. Statistical tests such as Friedman's test and Nemenyi post hoc analysis are carried out to ensure the statistical significance of the difference amongst models. The best performing IG prediction models are explained using SHAP values. This process not only highlights which features and model combinations perform well but also allows readers to design new features from existing features. This also shows the feasibility of the utility of ML pipelines in predicting IG values with known uncertainty and error distributions.

Chapter 5 Sleep Features for Prediction of Interstitial Glucose designs novel features for predicting IG values. This chapter marks a shift from comparing models to comparing feature combinations for predicting IG values. Recognizing the need for explainable and transparent models, sleep features and their importance can serve as a baseline for estimating the effect of sleep quality on glucose metabolism. This includes exploring the impact of sleep quality and durations on IG value prediction.

Finally, Chapter 6 serves as a synthesis, bringing together the research findings and their implications in the field of smart watches for glucose prediction. This Chapter critically analyses the conclusions derived from the research, providing a comprehensive discussion that situates the entire thesis within a broader context. Additionally, it highlights potential avenues for future research and underscores the current limitations of the thesis. The Chapter demonstrates the ongoing advancements in using smart watches for glucose prediction.

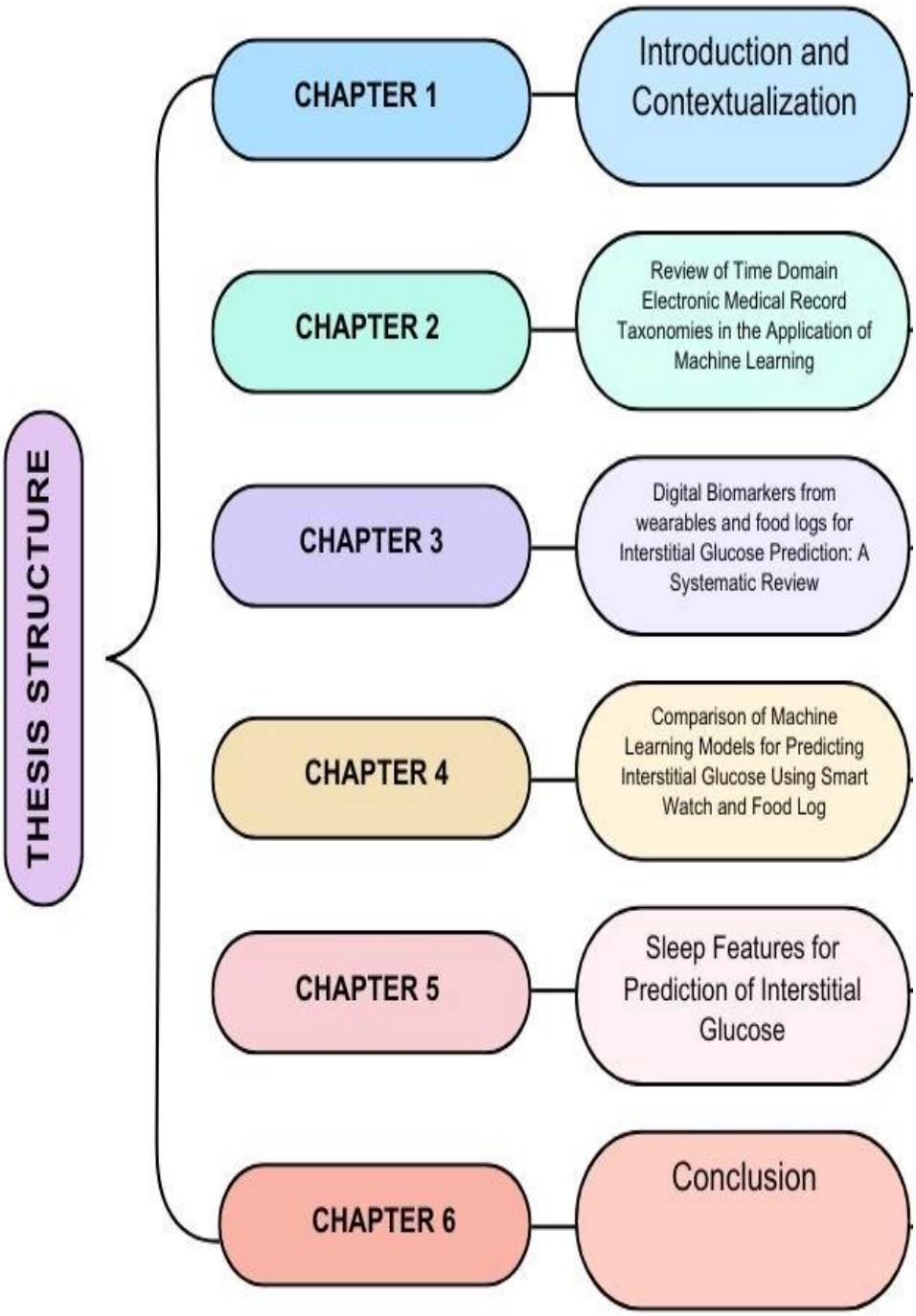


Figure 1.1: Organization of thesis into different chapters and titles.

2 Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning

2.1 Preface

The content of this Chapter is a copy of the article "*Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning*", published in peer-reviewed journal Electronics, available online.

Time domain data from healthcare domains presents unique challenges due to their sequential nature compared to other sensor data. Despite the variety across sensors in healthcare applications, most data are recorded as time series data. This Chapter provides a comprehensive review of different knowledge paradigms that exist in developing solutions that use time series data for machine learning applications to perform downstream tasks, examining essential requirements and variations based on diseases. This knowledge presents a broader overview of how time series data is handled in healthcare settings. The focus on time series data is derived from research questions 2 and 3, as answering these necessitate the use of smart watch sensors, which primarily record data in the form of a time series. Thus, awareness of pipelines used in applications that utilize machine learning and time domain healthcare data informs the research about how this data from smart watches should be transformed to predict interstitial glucose (IG) values. After this broader review, a more targeted literature review is carried out in Chapter 3 which highlights the data processing methods employed for measuring biomarkers from smart watch sensors (which primarily record data in the form of time series) that are used in predicting glucose levels. This review sets the stage for the next more targeted review to highlight the variety of processing techniques and methods used in measuring smart watch digital biomarkers that can be used to predict IG values measured. The smart watch data is also in the form of time series.

The review highlights safety, interoperability, and applications as crucial factors in the successful application of ML in time domain electronic medical record (EMR) data. Notably, there is a discernible trend towards developing inter-operability and explainability techniques as well developing new digital biomarkers from time domain EMR data.

2.2 Abstract

Electronic Medical Records (EMRs) help in identifying disease archetypes and progression. A very important part of EMRs is the presence of time domain data because these help with identifying trends and monitoring changes through time. Most time-series data come from smart watch devices monitoring real-time health trends. This review focuses on the time-series data needed to construct complete EMRs by identifying paradigms that fall within the scope of the application of artificial intelligence (AI) based on time series data. (1) Background: The question addressed in this study is: What are the taxonomies present in the field of the application of machine learning on time domain EMRs? This question helps with identifying various processes involved with applying AI on time series data like smart watch data that can be used for predicting interstitial glucose (IG) levels (2) Methods: Scientific databases including Scopus, Web of Science, and PubMed were searched for relevant records. The records were then filtered based on a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) review process. The taxonomies were then identified after reviewing the selected documents; (3) Results: A total of five main topics were identified, and the subheadings are discussed in this review; (4) Conclusions: Each aspect of the medical data pipeline needs constant collaboration and update for the proposed solutions to be useful and adaptable in real-world scenarios. While most data are different and the tasks performed are different, the broad steps used are Problem definition, collection of data, preprocessing, AI model application, and validation.

2.3 Introduction

Translational medicine (TM) includes collaboration between clinicians and scientists to develop artificial intelligence (AI) models that account for differing data sources, differing data collection methods, and other real-world factors. In some cases, TM reduces the time from development to deployment allowing effective communication (Baxi et al., 2022) between different stakeholders for shared goals. It is characterized by integrating the digital biomarkers, multi-omics profiling, model-based data, AI, biomarker-guided trial designs, and patient-centric companion diagnostics (Ahmed, 2022). Therefore, the taxonomies identified in this review are guided by translational Medicine (Hartl et al., 2021). (Jordan, 2015) presents the following components of the complete medical record: Connected fitness devices, patient-focused personal health records, individual behavioural patterns, pharmacy-focused medical adherence data., provider-focused medical records, connected medical devices, and genomic information. EMRs help in disease monitoring, pandemic monitoring, adjustment of lifestyles, hospitals, intensive care units, and integration of healthcare services.

In these complete electronic medical records (EMRs), the use of time series data is essential because most of the biomarkers are tracked as trends in time (Ewusie et al., 2020). This time series data is high dimensional and is sampled at higher frequency and thus requires statistical methods like Machine Learning (ML) to find patterns within it. ML is "The ability of computers to advise decisions based on the available data" (Ahmad et al., 2018). ML has been used in applied clinical studies for some time now (Baum, 1988), and recent renewed interest has been driven by increased data availability (Paganelli et al., 2022) and an increase in computational capacity (J. X. Chen, 2016). The most common medical data types are images, such as computerized tomography (CT) scans, time series data as well as tabular data such as the blood, urine and metabolic panels. Some noteworthy literature reviews cover computer vision techniques used to assist clinical decision making (Kumar & Mishra, 2020; Singh et al., 2020; Taghanaki et al., 2021), however, there is a need to review the wider research landscape relating to the application of ML in time series data in EMRs to give the relevant taxonomies, patterns in literature, emerging trends, and knowledge streams in this field.

Our review of the relevant studies has found numerous relevant publications including previous work by (Davy et al., 2015) that investigates the effectiveness of chronic care AI models employed in primary healthcare. More recent work by (X. Chen & Sun, 2022) reviews the use of probabilistic ML models applied to healthcare data while (F. Wang et al., 2020) reviewed the latest advancements in graph-based analytics in healthcare. The application of telemedicine in maintaining EMRs has recently been reviewed by (D. Gu et al., 2019) using a cite space analysis. Another related work is the scientometric review of the application of latent discriminant analysis in healthcare data undertaken by (Tran et al., 2019) and the emerging challenges when using unstructured EMRs have been deliberated by (Adnan et al., 2020) in the context of using unstructured big data in healthcare. Table 2.1 compares this work with earlier works and highlights the novel features of this research.

Our study has three novel features:

- It identifies taxonomies within the field after a systematic search of research databases.
- It finds these taxonomies based on the principles of translational medicine so that the reader may find all the information needed for a translational solution at one place.
- It identifies the core challenges and advancements in each taxonomy and provides a rigorous volume of literature to serve as a baseline.

Following are the research questions we will try to address:

1. What are the paradigms and taxonomies that fall under the umbrella of AI in time series data?
2. What are the latest advances in these domains?
3. What are the latest challenges in these taxonomies?

Table 2.1: Comparison with earlier works

References	Time Series	Disease Specific	Translational Medicine
(Davy et al., 2015)	✓	X	X
(Krishna et al., 2009)	✓	✓	X
(Bellamy et al., 2020)	X	✓	X
(Tran et al., 2019)	X	✓	X
(D. Gu et al., 2019)	✓	✓	X
This work	✓	X	✓

2.4 Methods

An in-depth literature search was performed to conduct the review, following the search strategy of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Takkouche & Norman, 2011).

A paradigm is a collection of elements based on a common lexeme (McCarthy, 2005). Identifying the paradigms of fields interacting with each other is imperative to achieving overarching solutions. In this review, these paradigms are based on an industry perspective of TM (Hartl et al., 2021) and TM principles. The three major constituents of TM are developing treatments and interventions, testing the proposed interventions' effectiveness, and deploying these applications in the real world (Adithan, 2017). In this review, all the paradigms in the field considered are presented by Figure 2.2 and the challenges in applying AI-based solutions to TM will also be discussed.

In this paper the literature is searched from the following databases: PubMed, Scopus and Web of Science and the papers are selected based on the between 2015-2022. The search term is (("Physiological sensors" OR "Biomedical sensors" OR "Bio-medical Sensors") AND ("Machine Learning" OR "ML" OR "Artificial Intelligence" OR "AI" OR "Deep learning" OR "DL" OR "Reinforcement learning" OR "Electronic health records") NOT ("Security and Privacy") NOT ("images") NOT ("Robot")). The search was limited to 2015 to July,4 2022. Only articles were included. Proceedings of various conferences

were excluded from the search. The language of the articles selected is English. The number of articles after the search and selection are 160.

After the records are collected, they are thoroughly read to answer the following questions for each paper to extract information from the identified records:

- What is the type of Data used?
- What kind of Algorithm is used?
- What pre-processing methods are used?
- What post-processing methods are used?
- What data privacy standards are observed?
- What interoperability or fusion techniques are used?

By answering these questions in the form of a synthesis table, the top three topics are identified within each paradigm. But if the top three topics in a paradigm are repeated in any other paradigm, the next three topics are also discussed in the section to give a holistic overview of the topic. For example, the top three topics in the subtopic: time series data and structured data are the same therefore in the structured data section, the next three topics are also discussed. The tables given in each section give examples of illustrative works within each paradigm, each row describes a unique illustrative example.

The results of this literature search are then organized under topics and subtopics based on the principles of TM.

2.5 Results

Collecting the records through the process previously described, then reading their methods and results, and using the principles of TM, enabled the paradigms to be identified. The subsequent discussion is arranged as follows, firstly we identify the most commonly occurring topics in the records that we have collected within each paradigm, then the challenges within the paradigms are elaborated over.

Figure 2.1 represents the results of the systematic literature review which shows the number of articles filtered out at each step and Figure 2.2 gives the fishbone (Ishikawa) diagram of the paradigms identified in the resulting literature.

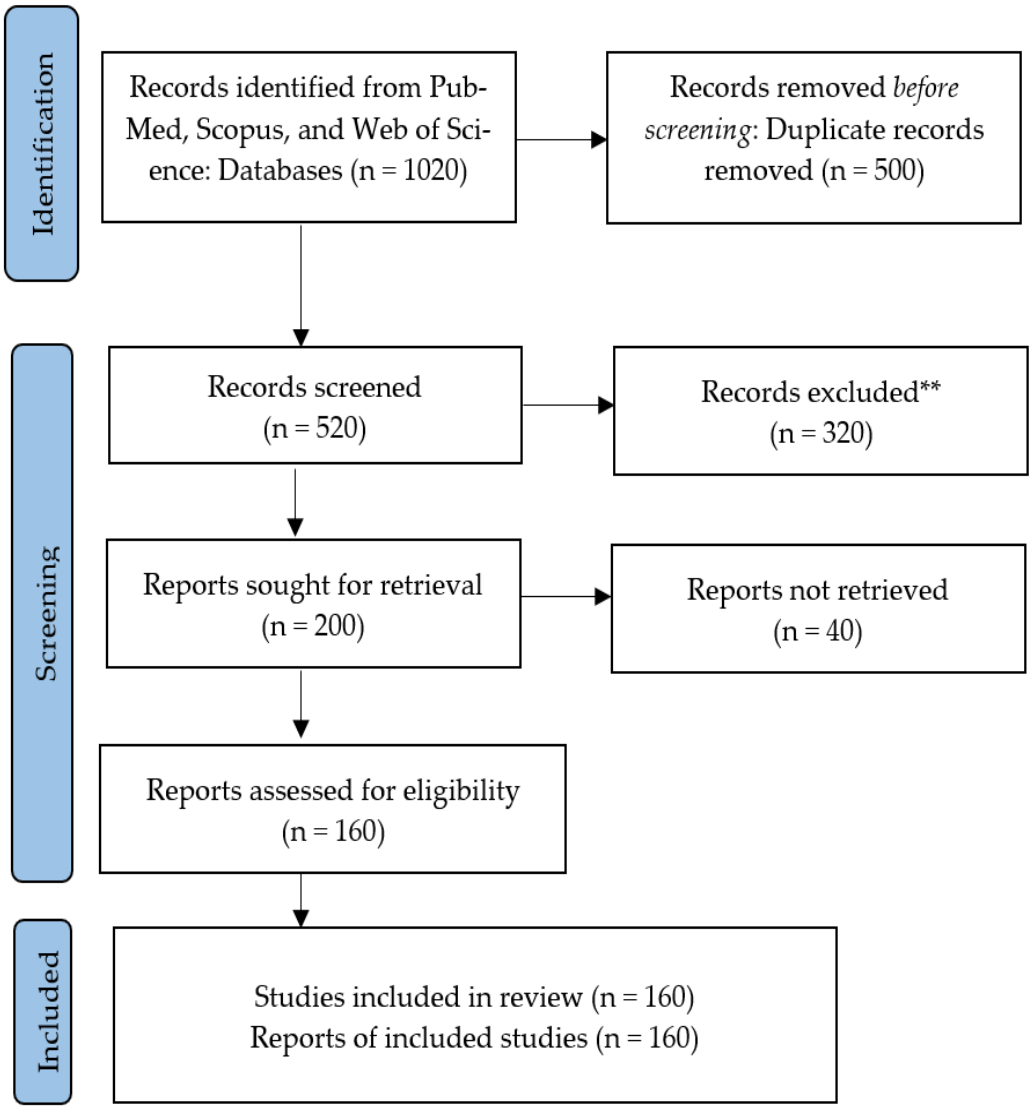


Figure 2.1: Literature search flow diagram

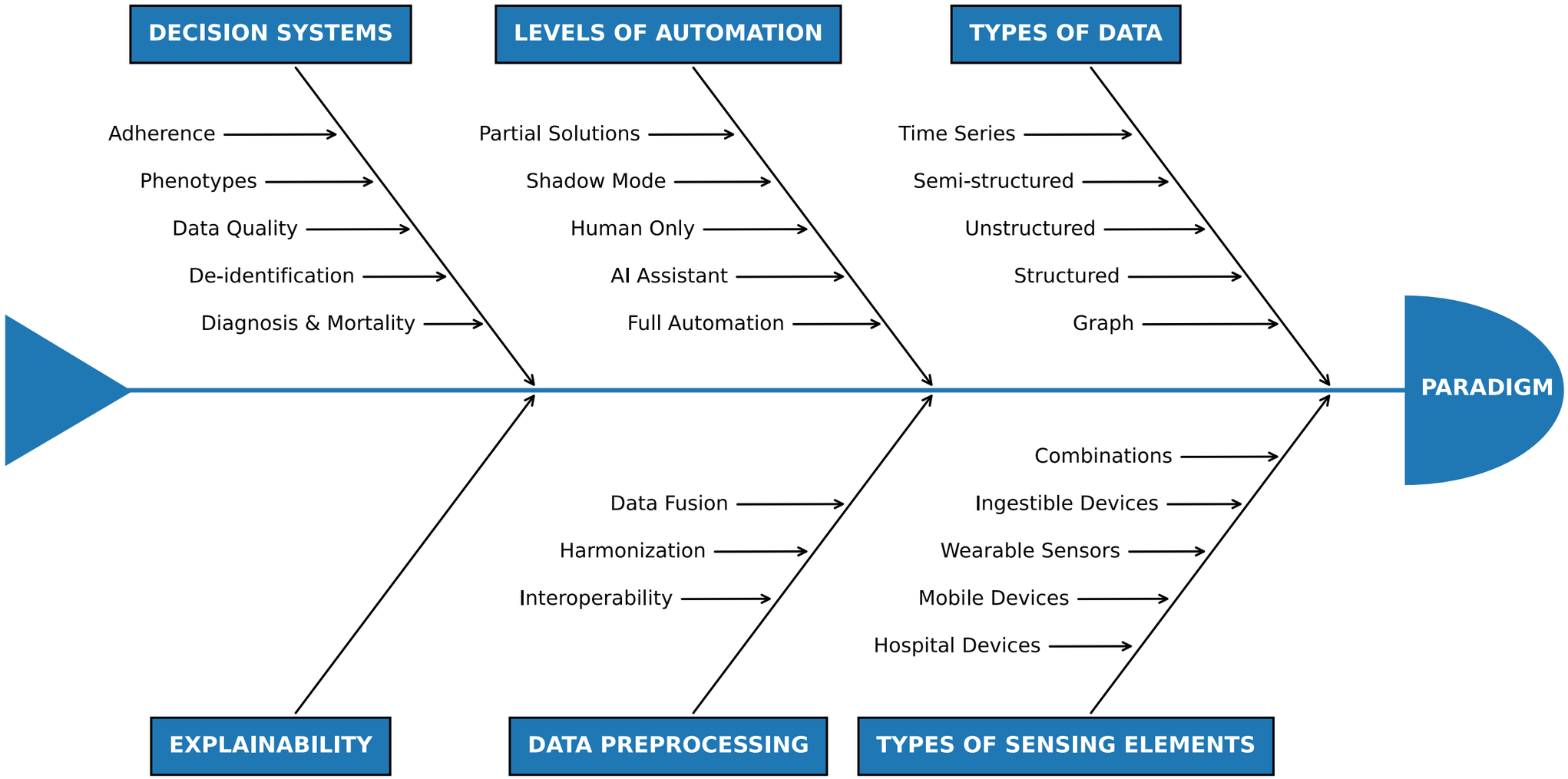


Figure 2.2: Ishikawa fishbone diagram of the identified paradigms from the literature selected for this review.

2.5.1 Representation of Data:

The way that the data is represented makes it suitable for the type of models to then be trained. The inductive biases of different ML models can be matched to different types of data representations. For example, as we will see in later parts of this work, the graph-based data representations are widely used for phenotyping (phenotype are the collection of observable effects of a disease), since the inductive bias of a graph-based ML model suits the relational nature of the problem of phenotyping.

Thus, the first axis along which we separated the data is the representation of data, which can be represented in a tabular form, or a time series, or a graph.

Time series (Tabular Representation): Time series data contains information from physiological events in the form of a time-varying biomarkers. Three leading solutions are specific to this data type: motif or pattern detection, data generation/imputation, and time series forecasting. Generative AI models overcome the lack of access to time series data by synthetically producing the missing and unknown data. Guided evolutionary networks (GENs) combine artificial neural networks and optimization algorithms such as genetic algorithms. They are used to fuse various information sources (Balasubramanian et al., 2016; S. Liu et al., 2020; F. Wang et al., 2020). GENs are also used to discover time-series motifs in ECG data (Pereira & Silveira, 2019). (Kaushik et al., 2020) Uses multilayer perceptron for time series forecasting in healthcare data.

Graph representation: Healthcare data is relational, hence, suitable for graph representation. Relational Data is characterized by the relations or dependence that exists amongst the rows and columns (Whitney, 1974). Graph-based techniques are used for developing graph-based representations of healthcare data, identifying clinical pathways and phenotypes of disease, and doing predictive modelling of disease and interventions. For example, (Choi et al., 2023; Q. Zhang et al., 2020) are some representative graph representations of healthcare data. (C. Liu et al., 2015) determines the temporal phenotypes based on graph representations of healthcare data. (Sood & Mahajan, 2017) is a fog-based temporal network graph analysis for the Chikungunya virus in India. (T. Wu et al., 2021) uses a proximity-preserving graph embedding to represent EMRs for hypertension. (Winter et al., 2003) incorporates metadata of the patients along with their vitals and lab results to learn a graph representation of electronic healthcare data. (Sharma & Bhatt, 2018) is a study that employs cryptographic techniques for information embedding in healthcare data. (Malik et al., 2020) is another knowledge graph-based phenotyping technique for subarachnoid haemorrhage. (Kalamaras et al., 2021) is a graph-based visualization for sensitive outcomes like

mortality or coma in medicine for healthcare data. (Q. Zhang et al., 2020) is a graph-based channel fusion for wrist pulse detection.

(Wright et al., 2019) uses graphs for learning a lower dimensional representation of drug-disease interaction. As illustrated in (Yue et al., 2020), graphs' main applications for medical interventions are drug-drug interaction, drug-disease interaction, protein-protein interaction, medical term classification, and protein function prediction.

The three main methods used in this paradigm are matrix factorization, random walk, and neural network-based methods. These include Laplacian methods as demonstrated in (Q. Zhang et al., 2020), Deep walk methods as shown in (Kulmanov et al., 2018) and Neural Networks as illustrated in (Finlayson et al., 2014). (Schrodt et al., 2020) is a literature review of graph theory applications on electronic healthcare data. It classified graphs into temporal data mining (C. Liu et al., 2015) causal and contextual (Kaur & Rani, 2015) and patient enteric graphs (Müller et al., 1996).

It is worth noting that there is no unique graph representation for sensor data or EMRs. Hence, most research focuses on developing graph-based presentations. One crucial research area is benchmarking and creating a numeric qualitative marker of adequate representation. Table 2.2 compares different graph-based solutions found in literature and their novel contributions are highlighted.

2.5.2 Structure of Data:

Another way to classify the type of healthcare data is the structure of available data. Most healthcare data is not structured against a set of rules. The structure of data dictates the kind of preprocessing required in the pipeline or the kind of algorithms that can eventually be used.

Structured Data: Structured data follows a definite set of rules or schemes (Palanisamy & Thirunavukarasu, 2019). The main issues that are dealt with using ML and structured data are data generation, data fusion, pattern detection, privacy preservation and prediction of outcomes. Privacy preservation is guided by HIPAA rules (Moore & Frye, 2019). Generative algorithms are used extensively to impute the missing data in the structured datasets (Abedi et al., 2022; Choi et al., 2023). Data fusion is another typical application of ML for combining two different kinds of structured data (Klompas et al., 2011). Federated learning that trains the models based on data from various decentralized devices is used extensively for privacy preservation of healthcare data (Aminifar et al., 2019; Kanwal et al., 2021; Wibawa et al., 2022; R. Xu et al., 2019). ML and structured data are also valuable in predicting the outcomes of interventions, for example, (Wright et al., 2019) analyses the user's choice in the event of alerts from clinical decision systems for potential drug-drug interference. (Vest et al., 2017) uses structured and unstructured data to find the social determinants of health characterized

by social behaviour, demographic features, and environmental factors of medical status and health care access.

(W. Wang et al., 2020) is a systemic review of records from PubMed and Web of Science on the detection of strokes from structured data and found the leading keyword to be mortality and the most used algorithms to be Neural Networks, Support Vector Machines, and XGBoost (Richter & Khoshgoftaar, 2018). Another review looked at the statistical and predictive machine learning models for cancer risk. It found the cox model (Therneau & Grambsch, 2000). (Therneau & Grambsch, 2000) to be a commonly used algorithm for predicting disease onset based on the input features. (Gopinath et al., 2020) uses AI to auto-complete structured clinical records based on context. (Shao et al., 2019) is a model to detect probable cases of dementia using structured and unstructured data. It uses a Latent Dirichlet Algorithm (LDA) for feature extraction and a logistic regression model. The key issues of research for structured data in healthcare are detecting phenotypes from EMRs (Banda et al., 2018; Sung et al., 2020), privacy and encoding of information (Boxwala et al., 2011; S. Kim et al., 2017; Lantz, 2016; Marble et al., 2020), data harmonization from various sources (Marble et al., 2020) synthetic data generation for research (Chin-Cheong et al., 2019; Guan et al., 2019; Walonoski et al., 2018), fairness and bias in the structured data (I. Y. Chen et al., 2020).

Semi-Structured Data: Semi-structured EMRs have no specific structure, enabling categorical data, meta-data, and numerical data to be entered in any field. The key areas in application of ML in unstructured data is conversion to structured data, predictive modelling, and interoperability of different kinds of data sources. For example, an application of ML with unstructured data for predictive modelling is used (Aggarwal et al., 2018) to derive contextual information to generate a semi-structured data from EMRs.

Table 2.2: Comparison of Graph based solutions

References	Application	Techniques used	Data	Contributions	Predictive	Descriptive
(C. Liu et al., 2015)		Attention Models		10% greater than RNN in disease prediction and 3% improved areas under ROC	✓	✓
(Kalamaras et al., 2021)	Temporal Phenotyping	Developed a temporal graph and normalized using hinge loss.	MIMIC-III	Predicted congestive health failure with an 80% accuracy. The area under the curve for patient readmission increased over 50 % from the spectral clustering	✓	✓
(Sharma & Bhatt, 2018)	Graph representation	Note Binning using hashing and term extraction using NLP.	STRIDE	Developed term and concept mappings	X	✓
(Q. Zhang et al., 2020)	Feature fusion	Graph-based multichannel feature fusion	Pressure and Phot-electric Sensors	93.1% accuracy in predicting diabetes from pulse detection data.	X	X

(Makarova & Lagerev, 2020) is a method to allocate resources from the knowledge of semi-structured healthcare data. (Hong et al., 2019) uses HL7 standards to develop the interoperability of structured, semi-structured, and unstructured data to develop phenotypes for obesity. (Batra & Sachdeva, 2016) is another such system that uses openEHR to this end. (Yuan et al., 2016) detects autism from semi-structured and unstructured data using a combination of models with skip-gram.

Unstructured Data: Most EMRs are unstructured (Miled et al., 2020). Key areas of research for ML applications in unstructured data are conversion amongst the various kinds of data structure and predictive modelling. An example for the predictive modelling used unstructured data is used by (Geraci et al., 2017) that employs unstructured EMRs to phenotype depression in youth. LDA and other dimensionality reduction methods are used to get the hidden information between different kinds of data and then leverage it for predictive modelling (Goh et al., 2021; Z. Xu et al., 2020; M. Zhao et al., 2022; Zuo et al., 2021). Aprioi algorithms and other Bayesian methods are used to convert unstructured data to structured data (J.-C. Kim & Chung, 2019; Malik et al., 2020; Song et al., 2017) and in so doing these works can also combine with the structured data to make predictions (Boustani et al., 2020).

Table 2.3: Comparison of solutions using unstructured data

References	Application	Techniques used	Accuracy	With Structured Data
(Geraci et al., 2017)	Detection of clinical depression	NLP	Specificity of 97% and a sensitivity of 45%	X
(Goh et al., 2021)	Disease Prediction	LDA	AUC 0.94, sensitivity 0.87 and specificity 0.87	✓
(Song et al., 2017)	Asbestosis detection	Apriori algorithms	Developed a relation amongst different disease clusters	✓
(Lin et al., 2017)	HPV detection	NLP	Mean area under the ROC curve of 0.861	X

Another technique that is relevant to the conversion of unstructured data to structured data is distant supervision. Table 2.3 compares the solutions related to healthcare using unstructured data. Distant supervision is a method for labelling the data by utilizing the known structures of similar data (Ling et al., 2019; Wallace et al., 2016). Exploratory text analysis is also used for pattern analysis for predictive modelling in this (Bjarnadottir & Lucero, 2018; Lin et al., 2017). NLP techniques are extensively applied to unstructured data to detect disease onset. Data harmonization and standardization is also an essential topic of discussion in unstructured healthcare

2.5.3 Type of Sensing Element

Types of Data are dependent on types of sensing elements. There are many types of sensing elements. These include smart watch sensors, mobile device data, ingestible sensors, medical devices from hospitals, and a combination of all or some of the factors mentioned earlier

- **Smart watch Sensors:** Smart watch sensors bridge the gap between assessment and onset prediction. The data sources measure the biomarkers from the physiological signals in real-time making this a vital component of multi-omics profiling (Y.-K. Wang & Chen, 2023)
- **Mobile Devices:** Along with real-time monitoring using mobile sensors, mobile devices also allow for input from the user, making them helpful in tracking medical adherence (Sempionatto et al., 2021).
- **Ingestible Sensors:** Drug adherence (Chai et al., 2015) and monitoring (Weitschies et al., 2021) are some applications of ingestible sensors.
- **Medical Devices from hospitals**: More connected medical devices can enhance healthcare quality for people in the hospital (G. Li et al., 2021).
- **Combinations:** The combination of the sensors enables the Internet of Medical Devices (Muhammad et al., 2021)

2.5.4 Data Preprocessing

As we have seen previously, data can come from various sources and in various forms. For the successful application of ML, this data the data must be harmonized and standardized. Data Harmonization Standards and Intelligent Interoperability techniques are the two classes along this knowledge stream. Another axes to classify data preprocessing techniques is the data fusion methods: feature level, data level, and decision level fusion. One more way to organize the data prior to analysis is through pre-processing techniques. These include filtering, feature extraction, and NLP techniques.

Data Harmonization Standards: Data Harmonization is a preprocessing technique that prepares different kinds of data to become compatible with each other. It allows the AI to access a diversity of information amongst researchers and institutions (Lucas et al., 2020). Some standards are specific to medical cases they deal with (Batra & Sachdeva, 2016; Baxter & Lee, 2021). However, there exists a set of general standards to ensure interoperability. The most common general standards are Health Level 7 HL7, openEHR, and ISO/IEEE 11073 Personal Health Data (PHD) standards (Laleci & Dogac, 2009).

Intelligent Interoperability: Intelligent Interoperability is use of ML or other algorithms to combine the information from different data sources in EMRs. In Intelligent interoperability of healthcare components, AI or some other rule-based systems are used

to automatically draw the relevant information from the EMRs or sensors data. These systems use different algorithms to ensure the interoperability of various data sources. Table 2.4 elucidates different interoperability strategies. Although these systems allow for effective data communication while ensuring information integrity, one key issue is allowing for the encoding of categorical features so that the information is stored effectively.**Data Fusion:** A physiological event can be observed with the help of various sensors, each sensing a unique aspect of the physiological event. The system must fuse or combine information from different sensing elements for a holistic understanding of the event. This is done at multiple levels. Following industry 4.0 techniques, healthcare systems comprise these sensing elements that are spread across time and space (smart watch sensors, ambulances, and hospitals). Fusing information from multiple sensors would provide a holistic picture of healthcare, including detection, phenotyping, disease progression, and other related data-powered solutions.(Dautov et al., 2019) is a combination of different layers of data fusion in connected healthcare, from individual sensors to sense medical events, to a network of connected devices, and finally, fusing information amongst various institutions. (Miao et al., 2019) defined different levels of data fusion. These include signal level fusion, feature level fusion, and decision level fusion. Kalman Filtering is a popular statistics method for signal level fusion and is widely used in biomedical sensor networks. Weighted averages are also widely used to penalize sensors with more noise in a sensor network (Djenouri & Balasingham, 2009; Nathan & Jafari, 2017). Particle Filtering, amongst various other variants, is also used extensively for signal level fusion in sensor networks in healthcare (Brady et al., 2016).

Table 2.4: Comparison of interoperability techniques

Name	Properties	References
Blockchain technology	Focused on patients rather than healthcare providers. Data is linked to the patient, aggregated and then sensitive information such as allergies is published on the blockchain, ensuring privacy and data immutability.	(Gordon & Catalini, 2018; Jabbar et al., 2020)
Internet of Things	It employs the principles of the internet of things for data interoperability. It uses the protocols of Message Queuing Telemetry Transport (MQTT) to publish the relevant patient information.	(Pathak et al., 2018)

Dynamic Semantic Web services	It uses the dynamic semantic web to convert the data into the HL7 framework.	(Balakrishna & Thirumaran, 2020)
Cloud Based Interoperability	It uses cloud-based models, for example, amazon web services, Microsoft Azure and IBM Watson to convert it into an openEHR or HL7 standard.	(I. Y. Chen et al., 2020)

(McKeever et al., 2010) uses temporal evidence theory for signal level fusion for activity recognition. Feature Level fusion means each sensing element's relevant properties are calculated and fused. (Cai et al., 2020) calculates a linear combination of features to get a new feature. (Miao et al., 2019) is a weakly supervised program for feature-level fusion. Decision Level Fusion is a way to fuse decisions based on different information streams. There exist many such systems in the context of healthcare (C. Chen et al., 2015; Hossain & Muhammad, 2017). The critical issue in all these is developing a plastic nature of fusion techniques. A plastic fusion technique would be flexible to change with the emerging problem because different features or data may have other significance for each model.

2.5.5 Decision System

The nature of decision systems is specific to the problem they deal with. One axe along which the decision systems can be classified is problems they solve from a clinical standpoint as well as from an engineering standpoint. The decision systems identified from the records are *data quality, phenotyping, medication adherence, graph representation of data, detection of disease, and Mortality Prediction*.

Data Quality: The quality of data acquired in healthcare is essential to the credibility of the predicted outcomes. Data quality issues are hard to identify in data with varying structures, shapes, dimensions, and sources. The dimensions of data quality as elaborated by (Weiskopf & Weng, 2013) are completeness (whether the relevant information is present), correctness (is the data correct), concordance (is it relatable to other data sources), plausibility (is any element in the EHRs making sense in the presence of other evidence) and currency (meaning how old is the data). These solutions will help to identify data quality issues, log them, encode them in metadata for datasets, help develop exclusion criteria of data based on its quality, and record the number of such problems. (Fox et al.) is one such work that creates a framework to carry out all the tasks and uses probabilistic models to detect temporal stability and plausibility in biomedical data. It employs probabilistic change detection using Jensen-Shannon distance principles of statistical control of posterior beta distribution. (Sáez et al., 2015) uses probability distribution distance to the same end. (Puttkammer et al., 2016) is a

measure of completeness by flagging incomplete data sources using the Delphi method. It also measured the same data quality dimensions using patterns in the number of patients and compared them and established that Delphi method performs better. (Taggart, Liaw, & Yu, 2015) evaluates the data quality of radio frequency identification (RFID) in nine phases in healthcare systems.

Phenotypes: Phenotypes are the combination of observable traits of disease for an individual. The data from EMRs is a set of data points related to interventions and the change in the states measured in lab tests. The data helps align heterogeneous disease progression into temporal phenotypes. This allows data science techniques to find the relation between disease, symptoms, and interventions. These are also linked to mortality prediction, disease progression, and observation of medically complex phenotypes. Most temporal phenotype identification methods deploy clustering techniques. Phenotypes are also used to identify rare diseases (Li et al., 2019) (Jia et al., 2018; Morley et al., 2014). These methods are rules-based (Morley et al., 2014) and graph theory-based (Ash & Rapp, 2014)

One of the critical challenges in AI-based phenotype is the representation of data. The data is being presented to domain experts but developing a metric that identifies the visual tools' efficacy to represent the temporal phenotypes is worthwhile. For example, in encoding information in edges and nodes of a graph, Silhouette diagrams (Lee, Rashbass, & Van Der Schaar, 2020) are more rich in information compared to graphs. *Category Theory* techniques can also be used for Phenotype identification. Category Theory provides a framework for phenotype identification by representing phenotypes as objects and their biological relationships, such as genetic or environmental influences, as directed arrows in a graph. The morphisms (arrows) capture functional or causal links between phenotypes, allowing for a structured view of how these traits are connected. By applying the concept of composition, where relationships between objects can be combined, Category Theory helps model the complex interactions underlying phenotypic traits, offering a mathematical approach to identifying and analysing phenotypes, however very little focused work in this domain comes from EMRs.

Deidentification: Automatic De-identification of EMRs is an active area of research. For this work blockchain is widely used (Mayer, da Costa, & Righi, 2020; Shi et al., 2020). (Kim, Heider, & Meystre) compares deep learning, rule-based systems, and shallow learning for de-identifying EMRs and argues that stacked learning is the most efficient. (Ahmed, Al Aziz, & Mohammed, 2020) deploys self-attention networks and stacked recurrent neural networks to de-identify the medical records. The main de-identification methods are neural networks, blockchain technology, and rule-based systems (Ahmed, Al Aziz, & Mohammed, 2020). Some Internet of Medical Things (IoMT) schemes uses

IoT protocols to preserve privacy while ensuring that critical information is relayed to the relevant stakeholder (Z. Guan et al., 2019).

Challenges in this field are the interplay of structured, unstructured, and semi-structured data. The data comes from various sources and categories and, in the case of categorical features with other features, must be collated before solutions can be designed.

Adherence: Adherence to suggested and prescribed medical regimens is a crucial component of healthcare. Healthcare is an integrated process; hence adherence is monitored by different sensing and AI techniques to ensure the efficacy of the interventions.

The key challenges in this domain are access to relevant data as the disease progresses. Here the importance of different features coming from the same sensors and additional sensors can change as the condition changes its phase.

Diagnosis and Mortality prediction: Disease prediction can help speed up the process of health care. In the case of critical systems, the idea of mortality prediction and their interplay with demographic information and phenotype can help save lives. It can also help in understanding the progression of the disease and can direct healthcare resources in the right direction.

(M. Chen et al., 2017) is a process for disease prediction using EMRs. It uses convolutional neural networks to this end. (S. Mohan et al., 2019) uses hybrid machine learning techniques to predict cardiovascular diseases. It uses a combination of random forest (RF) and linear classification models. (Venkatesh et al., 2019) develops a naive bayes model for disease prediction using EMRs. Table 2.5 represents the various AI Methods used for disease prediction.

Mortality prediction has also been carried out using ML from time domain EMRs for some time (Cooper et al., 1997). Mortality predictions have varying significance for different phenotypes (Rose, 2013; van Doorn, Stassen, et al., 2021). ML models are used widely in brain injuries (Raj et al., 2019; Rau et al., 2018) for mortality prediction.

The critical challenges in disease and mortality prediction are the development of explainable ML models because the model makes assumptions about the state of the health and how that may affect the disease state as well as risk of mortality. Another crucial issue in this domain is the development of ethical frameworks and safeguards to ensure that only useful information is relayed to the patients. One approach to carry out this research could be similar to identifying not only clearly defined ethical considerations but also the moral dilemmas like those explained in (Awad et al., 2018) for self-driving cars.

2.5.6 Explainability

The ML models developed using EMR data are explained in terms of the importance of input sources. An example of such explainability technique is Deep Learning Important Features or DeepLIFT (J. Wang et al., 2021; Zuallaert et al., 2018). DeepLIFT combines the importance of a feature as features move through the layers of the neural network. Local Interpretable Model Agnostic Explanation or LIME introduced in (Ribeiro et al., 2016) is also widely used to explain ML models that use EMR time domain data (Salih et al., 2021; Visani et al., 2020).

The explainability is either intrinsic in the models such as decision trees (DT) or post hoc. The first is when the models are designed such that they can be explained, and the former explains the predictions of the trained model. These explanations can be local (meaning for a single prediction) or global (meaning for the overall model).

Table 2.5: Comparison of different decision support systems using time domain EMR for healthcare applications.

Name	Summary	Application	References
Conversational Robot	Chatbot used for drug adherence	Drug Adherence	(Abd-Alrazaq et al., 2020; Vaidyam et al., 2019)
Ethics	Deliberates over the ethical questions arising from the usage of AI in Norm Adherence	Ethics	(Campbell et al., 2016)
Lifestyle Modification	It uses a web app to help monitor adherence, lifestyle modifications for example in the case of cancer.	Drug Adherence	(Golshahi et al., 2015)
Medication Adherence	It uses Machine Learning to do binary classification of the medication adherence for Parkinson's disease patients.	Remote Monitoring	(Molugulu et al., 2016)
Exercise Adherence	Uses Machine learning models to estimate likelihood to adhere to a physical exercise regimen using accelerators and other data sources.	Predictive healthcare	(Bavan et al., 2019)
Medication Adherence	Uses Machine Learning Models to identify the likelihood of non-adherence to medication from EMRs.	Predictive healthcare	(L. Wang et al., 2020)
	Uses data from smart watch sensors to measure drug adherence for a specific disease.	Remote Monitoring	(Aldeer et al., 2018)
	Uses cloud-based applications for the medication adherence in home hospitalizations	Remote Monitoring	(Chai et al., 2021)

These explanations not only rely on the input features but also the models, for example, partial dependence plot (which model feature interactions learned by a model) of a linear model such as linear regression looks like diagonally increasing or decreasing gradients as the model can only learn linear relationships. Thus, explanations alone, without any reference to the performance cannot be reliable estimates of the health outcomes predicted.

One example of intrinsic explanations in neural networks are attention mechanisms as they find the relevant neurons or the dataset components that are the most pertinent information used by the model. These mechanisms are at play in explainability models such as DeepSOFA, DeepHINT, and Grad-CAM (Hartono, 2020; Park et al., 2022; Shickel et al., 2019).

An example of classical explainable model is Least Absolute Shrinkage and Selection Operator (LASSO). LASSO can also be used as dimensionality reduction technique to explain the outcomes of a neural network. They are also used to describe healthcare outcomes from time domain EMR data (Bernardini et al., 2019).

Some explainability techniques draw the rules from the networks and such systems are also applied in healthcare (Ming et al., 2018; Xiao et al., 2016). Deep Taylor Decomposition is one explainability technique used in such systems (Montavon et al., 2017). SHAP values explain the outcomes of the model in terms of the input using game theory concepts and can explain the models robustly, and are used widely in time domain EMR data (Xie et al., 2022). However there still needs to be a lot more research in developing semantic explanations. Semantic explanations interpret and communicate model behaviour using high-level, human-understandable concepts instead of relying on low-level features like raw pixels or numerical values. With the advent of large language models (LLMs) and knowledge graphs, ontologies explaining the outcomes of a model can serve as a foundational step in semantic explanations of ML models using EMRs (Quintero-Narvaez & Monroy, 2024).

2.5.7 Levels of Automation

Another axis along which studies utilizing ML and time domain EMRs can be classified is levels of automation. Levels of automation for the health care studies using EMR time series data are discussed below:

1- Human Only: In this no AI is involved. For example the calculation of muscle atrophy using electromyogram (EMG) signals (Silva et al., 2018). This however in-volves the signal processing techniques for the representation of data.

2- Shadow Mode: In the shadow mode, the data between the interaction of the medical practitioner and other sources are logged, and the data is labelled

using the judgment of a qualified physician. This data is used to train a machine learning or an optimization algorithm. One such system developed by the ICL team as a reinforcement learning framework optimizing interventions retrospectively that al-lows a regulatory compliant pathway to clinical testing. This technique is used for sepsis treatment in the ICU (L. Li et al., 2020)

3- AI Assistant: It provides the physician with suggestions. Some systems use these to detect cancers; for example, one such system uses biomedical images and structured data to detect hepatocellular carcinoma in the AI assistant model (Menegotto et al., 2021).

4- Partial Solutions: Based on the data, the AI comes up with a diagnosis independently, but needs a physician's input

5- Full Automation: All the tasks in healthcare are provided by AI alone.

2.6 Conclusion

This review presents different paradigms in literature discussing application of AI in times series and graph-based healthcare data that is driven by TM. It looks at the complete pipeline, starting from data collection, harmonization, and quality dimensions. The decision systems are deliberated over, including various kinds of phenotyping, mortality detection, and other methods. We looked at the components related to the data, classifying them into multiple axes. We looked at the details related to explainability, algorithms and levels of automation.

In this review, we looked at the recent advances and state of the art in the various lexemes of the paradigms above. Data can be classified along multiple axes, including structure, source, and dimension. Most healthcare data is unstructured, which has been used in conjunction with structured data to predict outcomes related to healthcare. The data preprocessing techniques can help combine different types of da-ta, denoising and harmonizing to increase the reusability.

Blockchain has been used increasingly for the deidentification of data. Devices and sensors need to work interconnected with different data fusions, and this review has elaborated on different interoperability methods. Different decision systems and the algorithms that power them have also been elaborated on in this review.

The most recent works and reviewed which focus more on applying these solutions in the real world. Phenotype classification for the disease has been increasingly conducted using graph-based techniques. NLP is employed to infer valuable insights from unstructured healthcare data. Data is fused amongst different sources using various statistical and ML-powered techniques. Data generation has filled the gap of unavailable data using the power of generative AI.

The future work may include the identification of ethical dilemmas in healthcare interventions and personalized healthcare: continuous healthcare monitoring and better intervention methods.

There are many challenges associated with healthcare data collection for the so-called disease X (Higgins, 2021; Simpson et al., 2020).The more evolved diseases can be stopped from progressing in real-time using Mutliomics profiling and outlier detection (L. Li et al., 2020) Another challenge dealing with data derived from time-based sensor data is the integration of advancements in real-time systems. To this end, translational medicine is already defining some solutions. Another major challenge is generating data for groups for which this data is unavailable using generative AI

2.7 Chapter Summary

This chapter explores the growth in the field of ML applications for time series data in EMRs. However, different processing and filtering techniques need to be applied to the data to ensure interoperability, privacy, cleaning and making it suitable for subsequent downstream tasks. These processes can also be used in different healthcare settings, such as predicting glucose levels from smart watch data as it is also a time series data. This Chapter answers research question one of this work. Investigating time domain healthcare data and their ML applications is a compelling strategy to address the challenges for time domain healthcare data, leading to question two, which deals with investigating research landscape of digital biomarker design for IG prediction. Chapter 2 allows the development of the knowledge and context to carry out a narrower literature review that in Chapter 3.

3 Digital Biomarkers from Smart Watches and Food Logs for Interstitial Glucose Prediction: A Systematic Review.

3.1 Preface

The content of this Chapter is a copy of the article *"Digital Biomarkers from smart watches and food logs for Interstitial Glucose Prediction: A Systematic Review."*, submitted for peer-review journal of Computer Methods and Programs in Biomedicine.

The current state of healthcare research involving Machine Learning (ML) relies on finding representations that can help ML models detect and learn the patterns within them as discussed in Chapter 2. Such representations have been recently formalised as *digital biomarkers* defined as quantified metrics of the health state measured using digital devices. After the broad literature review in Chapter 2, there is a need for a more targeted review of the state of knowledge in the development of biomarkers from smart watch devices descriptive of glucose levels. For glucose prediction studies that use smart watch sensors, the utility of such digital biomarkers requires processing data through various steps involving windowing, filtering, cleaning, aligning, and performing statistical operations. This work compares other reported research that has converted smart watch sensors into digital biomarkers, highlighting the gaps in their design and model efficacy explored in more detail in Chapters 4 and 5.

These digital biomarkers are engineered and leveraged by ML models to make predictions. Hence, this review not only helps with answering research question 2 (which is about the research landscape of biomarker design for interstitial glucose (IG) prediction) but also 3 and 4, which are about comparing the performance of models for IG prediction and designing novel sleep related biomarkers for IG prediction. This chapter helps identify the gaps in the current state of research, such as sleep features and a lack of systematic comparison of IG prediction models that are explained in Chapters 4 and 5, respectively.

3.2 Abstract

Glucose monitoring is critical for managing metabolic health, traditionally done using Glycated Haemoglobin (HbA1C) tests and continuous glucose monitors (CGMs), which are invasive and expensive. Recent research has shifted toward predicting glucose levels using smartwatches and food logs via machine learning (ML), which relies on digital biomarkers extracted from the data. This review focuses on identifying these biomarkers and exploring the processes for converting data into useful markers. We systematically reviewed literature from PubMed, Scopus, ScienceDirect, and IEEE

Access (July 2014-February 2024), identifying 15 studies that used 100 digital biomarkers to predict 30 glucose-related markers. We adopted the recently proposed DACIA framework (Data acquisition, Aggregation, Contextualizing, Integration, and Action) to standardize the extraction of information from identified literature and classify biomarkers into the following categories: autonomic nervous system, physiology, food intake, activity, and circadian rhythms. DACIA framework proposes guiding questions that we use to provide practical pathways to design effective biomarkers predictive of IG levels using smart watch sensors and food logs. Our findings demonstrate that smartwatches and CGM data can be used to train ML models for non-invasive glucose prediction, offering a cost-effective alternative for metabolic health monitoring. However, challenges remain, such as the small number of studies and inconsistent methodologies. Future research should focus on standardizing methods and validating models across diverse populations to enhance their generalizability and utility. However, individualized models can still relate the activity and lifestyle measured using smart watches to the glucose levels, opening the door for more personalized care.

3.3 Introduction

Metabolic syndromes (MS), such as prediabetes, and metabolic diseases (MD) like diabetes pose a significant global health challenge. According to the World Health Organization (WHO), over 422 million people worldwide are affected by diabetes, a number projected to rise to 578 million by 2030 (Poolsup et al., 2013). MS is a collection of risk factors for MD, including hypertension, elevated glucose levels, and obesity (Aguilar et al., 2015). Importantly, MS is reversible with lifestyle changes and medication, making early detection and monitoring glucose levels crucial. Despite being largely preventable and manageable through lifestyle changes and medication, early detection and monitoring of these conditions remain inadequate (Nicolaisen et al., 2023).

MS has been proven to be difficult to screen (M. M. Kim et al., 2022). For instance, prediabetes often goes undiagnosed, with only 17.4% of affected individuals aware of their condition due to ineffective screening tools (Zimmet et al., 2016). The costs and inconvenience associated with regular glucose monitoring may be in part responsible for the ineffective screening of prediabetes as it clinically defined by fasting plasma glucose levels between 100 and 125 mg/dL. Thus, awareness of glucose levels can aid in screening and managing prediabetes. For individuals that have progressed to diabetes, frequent glucose monitoring can reduce the risk of hyperglycaemia and hypoglycaemia, both of which can have permanent and potentially fatal impacts (Patell et al., 2017).

Despite the clear benefits of regular glucose monitoring for both screening and management, achieving frequent measurements is challenging due to limitations in current monitoring methods**.** Traditional glucose monitoring techniques, such as

measuring glycated haemoglobin (HbA1C) levels, are invasive and limit the frequency of monitoring because they require blood sampling (Bennett et al., 2007). Continuous glucose monitors (CGMs) offer more frequent data collection by using a sensor inserted under the skin to continuously measure glucose levels in the interstitial fluid, providing near real-time interstitial glucose (IG). However, CGMs present challenges, including high cost, invasiveness due to subcutaneous sensor insertion, and logistical issues like limited data storage capacity and the necessity for regular data logging (Fonda et al., 2016). These factors make widespread screening and continuous glucose monitoring impractical for the larger at-risk population.

Given these challenges in glucose monitoring, there is a pressing need for alternative, non-invasive, and cost-effective methods to monitor glucose levels and assess metabolic health (Adams & Nsugbe, 2021). Advances in smart watch technology provide a promising solution. Smart watch devices equipped with sensors—such as temperature sensors, Heart Rate (HR) monitors, accelerometers, and electrodermal activity (EDA) sensors—can collect physiological data that can be used to find *digital biomarkers* (Bent, Cho, Henriquez, et al., 2021). Digital biomarkers are objective, quantifiable physiological and behavioural data that are collected and measured by means of digital devices such as smart watch sensors, smartphones, or other connected tools. These biomarkers provide real-time insights into an individual's health status by capturing continuous data in everyday environments, outside of traditional clinical settings (M. M. Kim et al., 2022).

ML models can leverage these digital biomarkers to predict glucose markers, offering a non-invasive approach to glucose monitoring (Adams & Nsugbe, 2021). However, the methods for collecting, interpreting, analysing, and translating health data from smart watches into actionable insights vary widely. There is a lack of systematic approaches to guide this process, which hinders the effective utilization of smart watch data in metabolic health management (Bartolome & Prioleau, 2022; Langer et al., 2024).

To address this gap, recent work has proposed framework named: Data, Aggregation, Contextualization, Interpretation, and Action (DACIA) framework, which provides a structured approach to processing smart watch data for biomarker design (Daniore et al., 2024). Building upon these developments, this paper aims to provide a comprehensive review of digital biomarkers derived from smart watch devices used in predicting glucose markers through ML models and practical steps that can be used to develop these biomarkers from raw data.

In this review, we systematically analyse literature from scientific databases such as PubMed, Scopus, and Institute of Electrical and Electronics Engineers (IEEE) explore databases, identifying studies that utilize smart watch sensors to predict glucose markers using digital biomarkers. We classify these biomarkers into categories related to

movement, eating behaviour, autonomic nervous system activity, and physiological parameters. Furthermore, we synthesize the information using the DACIA framework to provide a structured understanding of how these biomarkers can be measured, using different signal processing methods.

3.3.1 Related Work

Similar reviews have been conducted to list various biomarkers in the context of various diseases. These works are listed here, to highlight the novelty of this review. For example, (Wolkowicz et al., 2020) list biomarkers of glycaemic control for type 1 diabetes but do not include digital biomarkers measured using smart watches. (Keshet et al., 2023) review digital and smart watch devices for metabolic and cardiovascular diseases but do not focus specifically on digital biomarkers derived from smart watches for glucose prediction. Similarly,(van den Brink et al., 2021) discuss digital biomarkers for resilience but do not systematically categorize or compare them in the context of glucose monitoring. Chapter 3 is compared to earlier works in Table 3.1

Chapter 3 makes the following novel contributions:

1- We identify and compile studies that employ smart watch sensors for predicting glucose markers using digital biomarkers.
2- We provide a detailed list of digital biomarkers measured from smart watches that are used to predict glucose markers.
3- We classify these digital biomarkers into distinct categories to facilitate targeted research and application.
4- We apply the DACIA framework to synthesize and interpret the collected data, enhancing the systematic understanding of the field.
5- We provide practical steps that can be taken to measure digital biomarkers from raw data.

Table 3.1: Novel contributions of this work in comparison to similar works

Work	Smart watches	Glucose Markers	Digital Biomarkers	Systematic Synthesis
(Wolkowicz et al., 2020)	×	✓	✓	×
(van den Brink et al., 2021)	×	×	✓	×
(Keshet et al., 2023)	✓	×	✓	×
This work	✓	✓	✓	✓

This chapter is structured as follows. First, we outline the methodology for literature search and selection criteria. Next, we present the identified digital biomarkers under their respective categories and list practical steps needed to develop these biomarkers

from raw data. We then discuss the implications of our findings in the context of metabolic health monitoring and ML applications. Finally, we offer recommendations for future research and conclude by highlighting the potential of smart watch technology and digital biomarkers in revolutionizing glucose monitoring and metabolic syndrome management.

The research question answered in this work is:

- What movement, autonomic nervous system (ANS), eating and physiology related biomarkers can be drawn from smart watch sensors used to predict glucose markers?
- What DACIA can be associated with digital biomarkers for each category?

Table 3.2 summarizes the significance of this work

Table 3.2: Statement of Significance

Heading	Summary
Problem or Issue	Traditional glucose monitoring methods, such as CGMs, are invasive, costly, and inaccessible for widespread use, necessitating alternative non-invasive solutions.
What is Already Known	Smart watch devices have shown promise in collecting physiological data that can be used as digital biomarkers to predict glucose levels.
What this Chapter Adds	This paper systematically identifies 100 digital biomarkers from smart watch devices, demonstrating their utility in glucose monitoring through ML models.

3.4 Methods

To identify digital biomarkers derived from smart watch devices used in predicting glucose markers through ML models, we conducted a systematic literature review (SLR) following established guidelines. The objective was to collect relevant studies, extract the digital biomarkers and models used, and synthesize this information to address our research questions.

We designed our search strategy based on the PICOS framework to ensure a comprehensive and systematic approach. The PICOS components are as follows:

- Population (P): Human participants using smart watch devices for glucose monitoring.
- Intervention (I): Not applicable, as our focus is on the identification of digital biomarkers
- Comparison (C): Not applicable, as our focus is on the identification of digital biomarkers

- Outcome (O): Measurement and prediction of IG markers using ML and artificial intelligence techniques for glucose marker prediction.
- Study Design (S): Studies employing predictive models using smart watch sensor data.

Based on these components, we formulated the following search terms for our query:

- ML Terms: "Machine Learning," "ML," "Artificial Intelligence," "AI"
- Smart watch Sensor Terms: "Smart watch Sensors," "Activity Trackers," "Wrist Worn Sensors," "Smart Watch"
- CGM Terms: "Interstitial Glucose," "IG," "Continuous Glucose Monitor," "CGM"

These terms were combined using Boolean operators to create the search query:

("Smart watch Sensors" OR "Activity Trackers" OR "Wrist Worn Sensors" OR "Smart Watch") AND

("Interstitial Glucose" OR "IG" OR "Continuous Glucose Monitor" OR "CGM")

We applied this search strategy to the following electronic databases to identify relevant studies published up to February 5, 2024:

- PubMed: 337 Articles
- Scopus: 5745 Articles
- IEEE Explore: 224 Articles
- Science Direct: 298

To ensure the relevance and quality of the studies included in our review, we established the following inclusion and exclusion criteria:

Inclusion Criteria:

- Original research articles.
- Published in English.
- Published between 2014 and 2024.
- Studies involving human participants.
- Studies using ML to predict glucose markers from smart watch sensor data.

Exclusion Criteria:

- Review articles, conference abstracts, editorials, and letters.
- Studies focusing on the design of new sensing methods or hardware, to maintain focus on existing smart watch technologies.
- Animal studies or in vitro experiments.

The study selection process followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. All records retrieved from the database searches were exported into a reference management software, and

duplicates were removed. Titles and abstracts of the remaining studies were screened independently by two reviewers (SM and HA) to assess eligibility based on the inclusion and exclusion criteria using Covidence review management website (Babineau, 2014). Full-text articles were obtained for studies that met the initial screening criteria. These articles were further assessed for eligibility. Studies that satisfied all criteria were included in the final review.

From the selected studies, we extracted the following information:

- Smart watch Devices Used: Type and specifications of the smart watch sensors employed.
- Digital Biomarkers Identified: List of digital biomarkers derived from smart watch data.

Table 3.3: Guiding questions to synthesize information for relevant digital biomarkers

Phase	Guiding Questions
Data	• Which smart watch sensor data is used to monitor relevant biomarkers? • Which other data types can meaningfully complement the smart watch sensor data for higher digital biomarker accuracy? • What could contribute to the missing data?
Aggregation	• What is the time window length suitable for the biomarker estimate? • Is the sampling frequency of sensor suitable for the desired window length? • What is the minimum wear time required to measure the effect of biomarker for relevant glucose marker prediction?
Contextualization	• Are the primary signals influenced by the timing of measurements (such as weekday or season) or by specific participant characteristics (like gender, age, or body mass index (BMI))?
Interpretation	• What degree of change in signals related to the outcome of interest would be deemed significant and clinically relevant when assessing digital biomarkers from smart watches with glucose markers?

Actions	• What are considered meaningful interventions in response to significant deviations in digital biomarkers?

- ML Models: Details of the ML models used for predicting glucose markers.
- Outcomes: Glucose markers predicted, and the performance metrics reported.
- Application of DACIA Framework: How each study addressed the Data, Aggregation, Contextualization, Interpretation, and Action phases.

To identify the DACIA attributes of the data the guiding questions are used for each biomarker type given in Table 3.3.

Furthermore, the ML models are compared based on various performance metrics. Models that categorize CGM values into different classes express their performance metrics using accuracy, which is defined as the percentage of correct predictions out of the total predictions, providing a measure of the overall effectiveness of the model in classifying CGM values correctly. Precision, which is the ratio of true positive predictions to the sum of true positive and false positive predictions, indicates the proportion of positive identifications that were correct. Recall, or sensitivity, is the ratio of true positive predictions to the sum of true positive and false negative predictions, measuring the model's ability to identify all relevant instances within a dataset. The F1 Score, the harmonic mean of precision and recall, provides a single metric that balances both concerns, especially useful when the class distribution is imbalanced. Finally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures the ability of the model to distinguish between classes, with a higher AUC indicating better model performance. By employing these metrics, researchers can thoroughly evaluate the effectiveness of their models in predicting CGM values and make informed decisions on potential improvements and adjustments.

Biomarker efficacy for predictive models can be compared based on feature importance scores, indicate how much each biomarker contributes to the model's predictive power, identifying the most influential biomarkers. The article screening process was based on PRISMA protocol and is given in Figure 3.1.

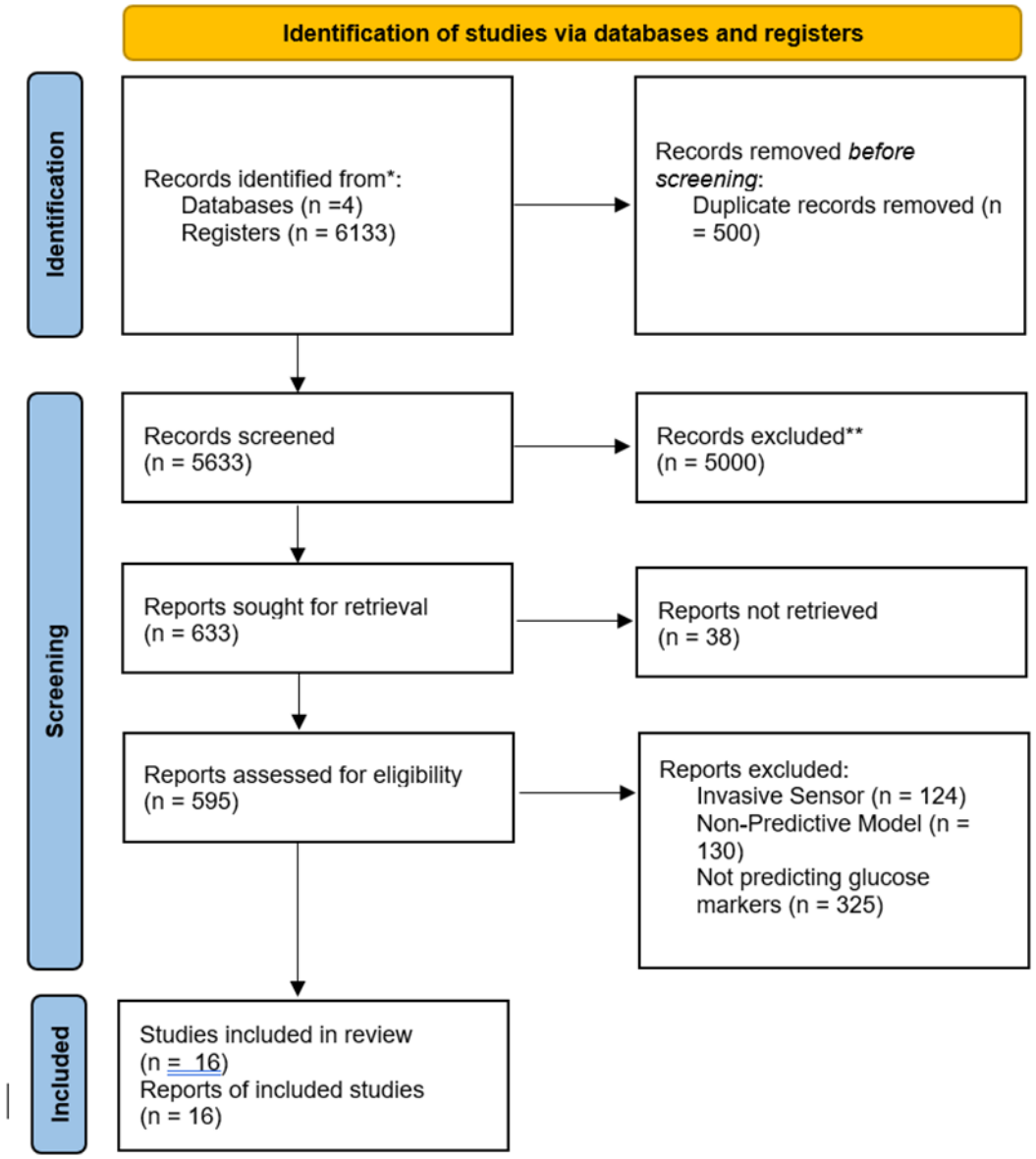


Figure 3.1: PRISMA diagram highlighting the number of articles removed at each selection step and the number of articles (16) used in this study.

3.5 Results

In this section, we present the digital biomarkers used in the identified studies from literature search to predict glucose levels and markers. These studies employ data from smart watch activity trackers, food logs, insulin pumps, and CGMs. We begin by summarizing the glucose markers predicted. Next, we discuss the digital biomarkers used by ML models to predict glucose markers. Finally, we provide an overview of the models employed and conclude with a discussion of our findings.

Figure 3.2 is a pie chart that shows the distribution of population types across the studies identified in this SLR showing that a significant proportion focused on individuals with T1DM, highlighting the importance of glucose monitoring in this group. T1DM patients rely entirely on exogenous insulin due to the autoimmune destruction of pancreatic beta

cells, leading to a complete lack of endogenous insulin production. This dependence results in rapid and unpredictable fluctuations in blood glucose levels, making them highly susceptible to acute complications such as hypoglycaemia and hyperglycaemia, hence explaining the higher portion of T1DM populations in the studies. Moreover, the widespread use of technologies like CGMs and insulin pumps among this group generates rich datasets that are ideal for ML applications.

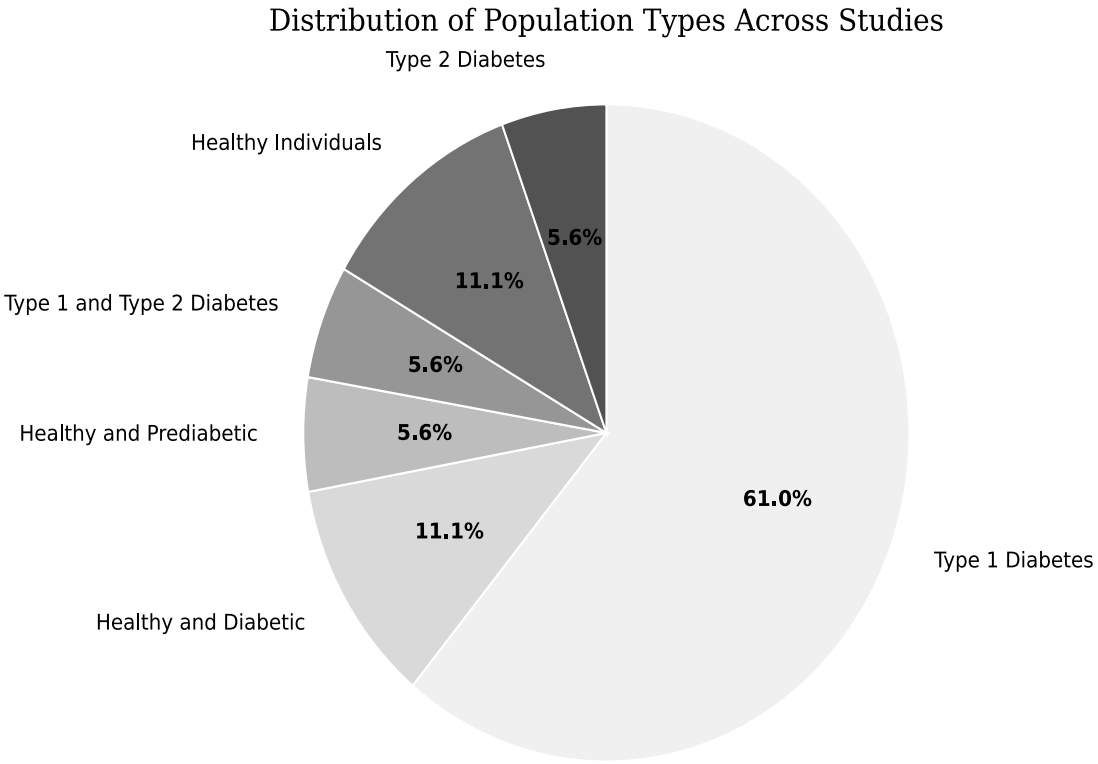


Figure 3.2: Types of populations used in glucose prediction studies reviewed in Chapter 3 highlight the prevalence of T1DM participants in these studies and public datasets.

As shown in Figure 3.3., CGMs are the most frequently utilized sensors across the reviewed studies, reflecting their pivotal role in diabetes management and research. The predominance of CGMs can be attributed to their ability to provide real-time, continuous glucose measurements, which are essential for both clinical decision-making and the development of predictive models. Smart watch sensors (accelerometers, HR monitors) and food data highlight the data sources used in ML models, predictive of IG levels. This trend represents the growing interest in personalized medicine and the utilization of multimodal data to improve diabetes care.

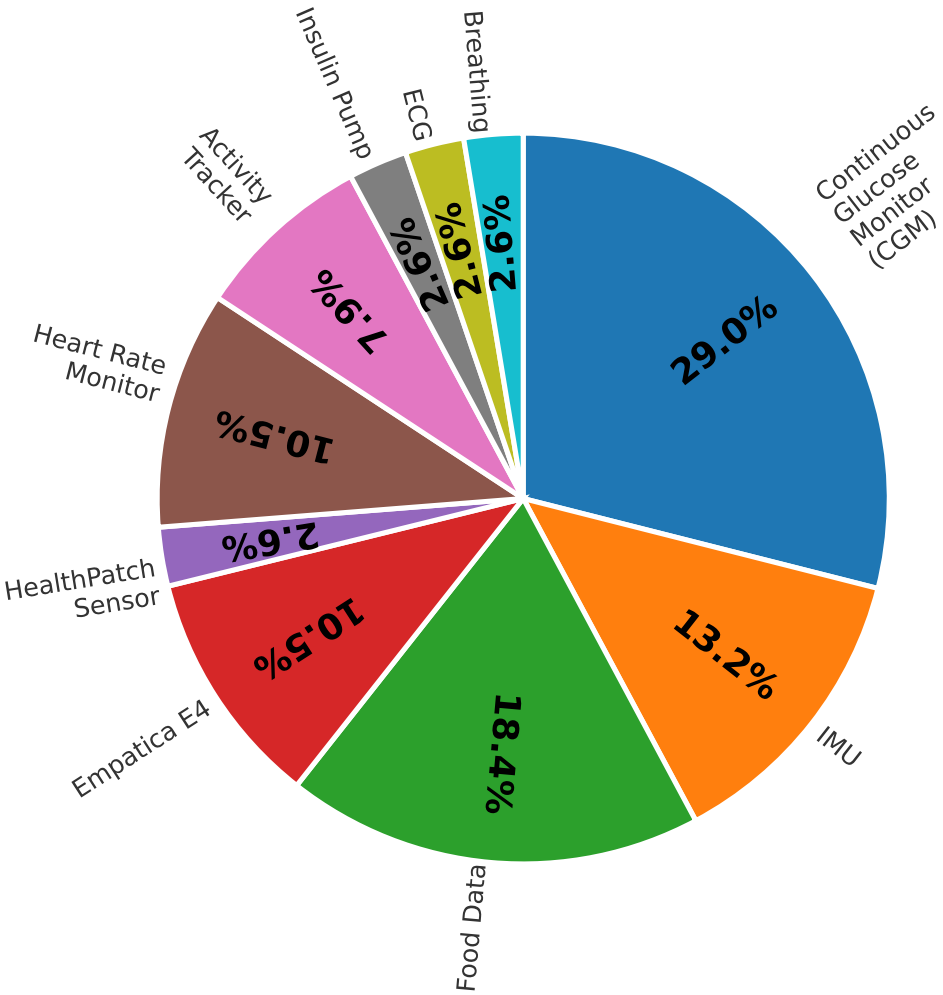


Figure 3.3: Types of Sensors Used Across Studies. This figure illustrates the distribution of different sensor types utilized in the reviewed studies

As depicted in Figure 3.4, there is a notable trend in ML models based on the predictive task. Regression models, particularly Neural Networks (NN) and Long Short-Term Memory (LSTM) networks, are predominantly utilized for forecasting IG values due to their ability to model complex temporal dependencies in the data. On the other hand, classification tasks, such as detecting hypo glycaemic events, frequently employ Support Vector Machines (SVM) and Gradient Boosted Decision Trees (GBDT), which are effective in handling high-dimensional data and distinguishing between different glycaemic states. This distribution highlights the importance of selecting appropriate ML models that align with the specific objectives of the study, ultimately contributing to advancements in personalized diabetes management.

The information from the selected literature is synthesized in each subsequent section

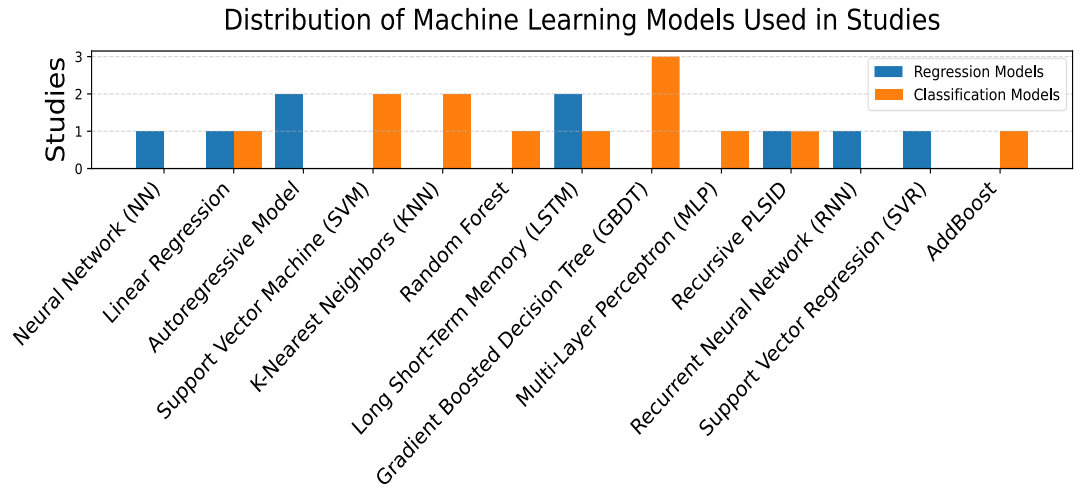


Figure 3.4: Distribution of ML Models Used in Studies

using the guiding questions outlined in Table 3.1 (refer to the Methods section). Each biomarker was classified into relevant categories, and the DACIA framework was applied to analyse the information for each class.

Firstly, we will summarize the digital biomarkers predicted by these studies and their roles in regulating levels of glucose, then we will discuss all the features that are used by the models to predict these biomarkers and finally we will discuss the models that are employed in these studies, concluding with a discussion on the findings.

3.5.1 Glucose Markers Predicted

The glucose markers predicted in these studies can be broadly categorized into continuous values (Bartolome & Prioleau, 2022), such as Time in Range (TIR), and categorical predictions, such as personalized classifications of high, low, and normal glucose levels (Bent, Cho, Henriquez, et al., 2021). The following glucose markers were predicted:

Diabetes Prediction: Some studies focused on predicting the presence of diabetes using data from smart watch sensors(Ganie et al., 2023; M. Li et al., 2020; Site et al., 2023). Early detection through non-invasive methods can facilitate timely interventions and management strategies.

Personalized high, low and normal glucose: In healthy and prediabetic individuals, glucose levels typically remain within a narrow range (100–125 mg/dL), making it challenging to detect significant fluctuations using standard thresholds, thus (Bent, Cho, Henriquez, et al., 2021) proposed personalized definition of high, low and normal glucose. Initially for the CGM values a mean, and standard deviation is calculated daily, all the values that stay within the standard deviation from the mean are considered personalized normal glucose levels. The values that exceed the standard deviation from mean are considered high and those that are standard deviation below the mean are considered low values.

TIR: TIR of glucose levels is defined as the time that the glucose levels measured with CGM stay within a defined range. The range is defined by clinical guidelines and is typically 70-180 mg/dL (V. Mohan et al., 2023). Some studies predict this value as one of the markers of glycaemic control (Bartolome & Prioleau, 2022; Bent, Cho, Wittmann, et al., 2021a).

Time Above Range: As the name suggests, time above range measures how much time the value of CGM stays above the normal limit (100-125 mg/dL) and is used as a ground truth for models of glycaemic control (Bartolome & Prioleau, 2022; Bent, Cho, Wittmann, et al., 2021a; van Doorn, Foreman, et al., 2021).

Bolus Insulin (units/day): It is a biomarker that can be measured using an insulin pump. It is defined as the units of insulin delivered by the insulin pump in response to rise in CGM values. The higher it is the lower the glycaemic control is. (Bartolome & Prioleau, 2022) has shown a combination of various biomarkers and their ranges using statistical tests that define good and bad glycaemic control and have shown that for TIR >70% defined as good control, the frequentist probability of 0-10 units a day is around 10%. This, however, is also dependent on the model of the insulin pump.

Basal Insulin (units/day): It is also measured using an insulin pump and is and depends on the pump. It is the amount of insulin injected by the insulin pump in fasting.

Statistical Biomarkers: 25, 50, 75th quarantines, inter day standard deviations, daily means, standard deviations, and various other statistical biomarkers are drawn and predicted from CGMs this however requires some more statistical tests to relate them to the control of IG.

ML models are used to predict these values as ground truth, but these values are predicted based on the features of the input data, in the following sections we would discuss the various input features to these models under the headings of movement related features, food related features, stress related features, and physiological features. Figure 3.5 captures the number of studies in this review for each marker.

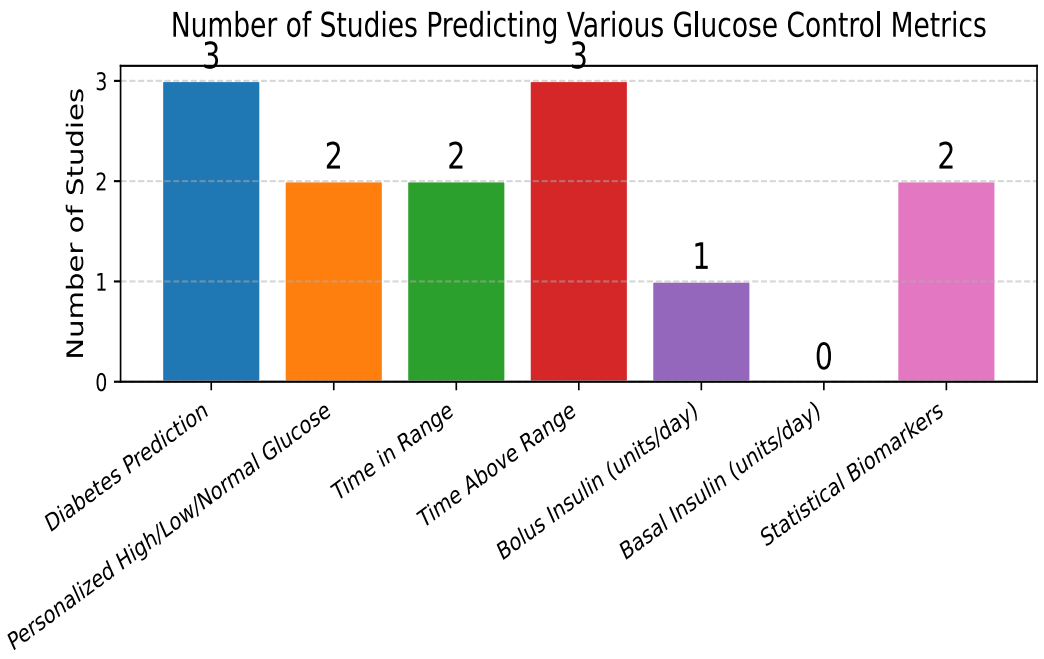


Figure 3.5: Number of studies predicting each glucose marker

3.5.2 Food related biomarkers

The quantity, quality, and timing of food intake significantly affect IG levels. Accurate prediction of glucose levels, therefore, necessitates detailed information about dietary intake. To capture this data, glucose prediction studies often require participants to log their consumed food using diaries or digital devices (Zahedani et al., 2023). Food-related digital biomarkers are then estimated by matching the logged food items with nutritional databases to determine their nutritional content (Bohn et al., 2022).

Compliance of food logs decreases with time (Turner-McGrievy et al., 2019) To enhance compliance, some studies have incorporated the use of food photographs, as taking pictures is generally more convenient for participants. Datasets such as OhioT1DM and D1NAMO include images of consumed food alongside CGM values. (B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021). However, converting food images into accurate nutritional information is more complex than processing traditional food logs(Khan et al., 2022).

There is a delayed effect of food on IG levels due to digestion and absorption processes. To capture this delayed effect, glucose prediction studies use rolling averages of nutrients in food over specific time windows. These time windows are typically set to 2, 4, and 8 hours. Rolling averages are calculated by summing the nutritional values of food consumed within the window and dividing by the number of data points, effectively smoothing out short-term fluctuations and highlighting trends. Table 3.4 provides the food biomarker definitions.

Table 3.4: Food related biomarkers

Feature	Definition	References
Estimated Carbohydrates	Estimated Carbohydrates entered an insulin pump for it to be able to calculate the bolus insulin	(Bartolome & Prioleau, 2022)
Calories	2-, 4- and 8-hour rolling averages of calories	(Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a)
Proteins	2-, 4- and 8-hour rolling averages of proteins	(Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a)
Sugars	2-, 4- and 8-hour rolling averages of sugars	(Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a)
Carbohydrates (from food database search)	2-, 4- and 8-hour rolling averages of carbs	(Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a)
Eat Counts	2,4,6,8,12, and 24 hour counts for eating events	(Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a)

To systematically analyse the food-related digital biomarkers, we applied DACIA framework. Table 3.5 presents the synthesis of information using the DACIA framework.

Table 3.5: DACIA information synthesis table for food biomarkers

Phase	Guiding Questions	Answers
Data	• Which smart watch sensor data is used to monitor relevant biomarkers? • Which other data types can meaningfully complement the smart watch sensor data for higher digital biomarker accuracy? • What could contribute to the missing data?	• Food biomarkers are measured using (Cho et al., 2023a) diaries, or food pictures (Marling & Bunescu, 2020). • To increase the accuracy of measurement of digital biomarkers, labelling or additional sensors are used in similar works (Sempionatto et al., 2021). (Karim et al., 2020) used a NN to predict glucose values. • The missing data in these biomarkers can come from irregular and incorrect labelling.
Aggregation	• What is the time window length suitable for the biomarker estimate? • Is the sampling frequency of sensor suitable for the desired window length?	• Works have used 2,4- and 8-hour windows to calculate these biomarkers using rolling means for the prediction windows. • Analysis for the effective sampling frequency needs to be analysed statistically.
Contextualization	• Are the primary signals influenced by the timing of measurements (such as weekday or season) or by specific participant characteristics (like gender, age, or body mass index (BMI))?	• Nutritional effect of the food depends on the weight and height of the participant (Dewettinck et al., 2008) but these works input the food values to search a database AFCD etc and used the results.
Interpretation	• What degree of change in signals related to the outcome of interest would be deemed significant and clinically relevant when assessing digital biomarkers from smart watches with glucose markers?	• The feature importances of the (Bent, Cho, Henriquez, et al., 2021) found that food features about 37% important in predicting glucose markers.
Actions	• What are considered meaningful interventions in response to significant deviations in digital biomarkers?	• Significant changes in food biomarkers can trigger hyper and hypoglycaemia.

3.5.2.1 Practical steps for calculating food biomarkers:

In the reviewed studies, food-related digital biomarkers are calculated by matching logged food items to entries in nutritional databases. The databases commonly utilized include: MenuGene (Pinter et al., 2011), Australian Food Composition Database (AFCD) (Melville et al., 2023), and Food Composition Table for Bangladesh (FCTB) (Shaheen et al., 2013). AFCD, MenuGene and FCTB provide detailed nutritional information, such as macronutrient and micronutrient content, for a wide variety of foods.

After retrieving the nutritional components from these databases, some studies input this information into absorption models to simulate the digestion and absorption processes. For instance (Karim et al., 2020) employed glucose absorption curves derived from the model proposed by (Arleth et al., 2000). The outputs of these absorption models are often smoothed using averaging filters to account for physiological delays and variability in nutrient absorption rates and can also be used as biomarkers for prediction of IG levels.

Despite these methodologies, there is a significant need to simplify the process of food logging and nutrient estimation. Manual food logging is time-consuming and can suffer from non-compliance or inaccuracies due to user error (Turner-McGrievy et al., 2019). To address these challenges, advancements in computer vision and AI can be used. AI models can then estimate the nutritional composition based on the visual information. The development of annotated datasets containing images of food labelled with the type of food and caloric information is a critical first step in enabling these technologies. Such datasets will facilitate the training and validation of the models, ultimately improving the accuracy and efficiency of food-related biomarker estimation.

3.5.2.2 Food Biomarkers Efficacy in Glucose Marker Prediction:

Food-related digital biomarkers have demonstrated significant efficacy in predicting glucose markers. For instance, Karim et al. (2020) compared future glucose values using glucose absorption curves from the Arleth et al. (2000) with those using only past CGM values and the results indicated that incorporating absorption-based biomarkers improved model performance by 21.3% in terms of MAE for CGM values. Similarly,(Bent, Cho, Henriquez, et al., 2021) conducted feature importance analyses in their glucose prediction models and found that food-related features accounted for approximately 37% of the predictive power, making them the most significant contributors compared to other features such as physical activity or physiological signals.

Furthermore, ablation studies performed by (Zahedani et al., 2023) reinforced the importance of food biomarkers. By systematically removing different feature sets from

their models, they observed a notable decline in predictive performance when food-related features were excluded.

Incorporating detailed dietary intake data, especially when processed through physiological absorption models, substantially enhances the accuracy of ML models predicting glucose levels. This highlights the need for continued research and development in simplifying and improving food logging methods, as well as integrating advanced modelling techniques to capture the complex relationships between food intake and glucose dynamics.

3.5.3 Activity biomarkers

Physical activity, defined as any bodily movement that results in energy expenditure, is a significant predictor of glucose markers. Like food consumption, activity biomarkers have a time-lagged effect on glucose levels and are also measured with rolling averages. Activity levels are typically measured using accelerometer data. A Micro-Electro-Mechanical System (MEMS) 3-axis accelerometer measures continuous forces along three spatial dimensions: x, y, and z. It detects changes in acceleration by measuring the displacement of a small mass within the sensor, influenced by gravitational and inertial forces.

To better understand the functioning of accelerometers, Figure 3.6 illustrates the axes of a 3-axis accelerometer and how resultant acceleration is calculated.

The preprocessing and transformation of accelerometer data into activity biomarkers depends on the type of smart watch device used. For example, devices such as the Empatica E4 output raw accelerometer data (B. Bent et al., 2021; Maritsch et al., 2020) whereas other smart watches such as Fitbit output step counts and flights of stairs climbed (Bertachi et al., 2020). There is a difference in the output ranges of the g values based on the sensors (for example, Empatica E4 outputs data in ranges ± 2g whereas Apple watch outputs the data in fractions of g) thus requiring efforts to align data coming from various sources.

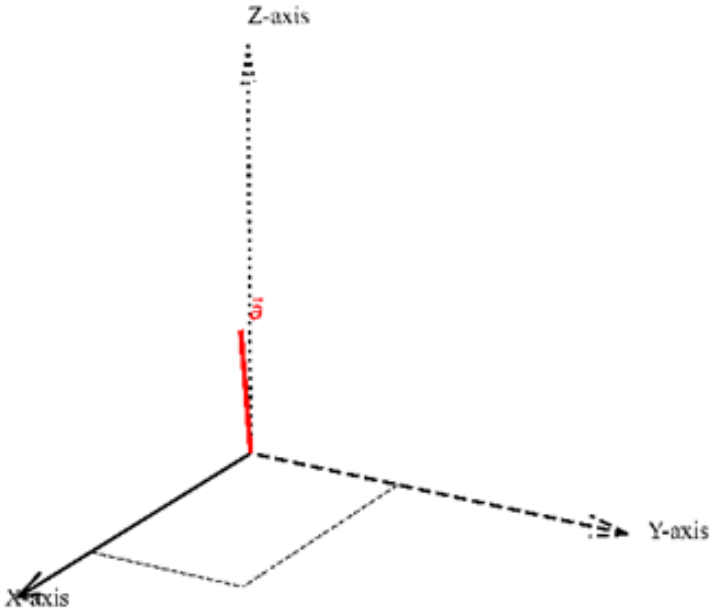


Figure 3.6: Representation of a 3-Axis Accelerometer with Acceleration Vector

This difference in data has its own merits and demerits. For example, raw accelerometer data requires post processing (B. Bent et al., 2021; W. Gu et al., 2017) whereas processed data such as step counts, or flights of stairs climbed limits the information richness(HAYERI, 2018; Rodríguez-Rodríguez, Rodríguez, et al., 2019).

In most works accelerometer gives three accelerations (a_x, a_y, a_z) relative to the part of body (in most cases wrist) it is attached to. The resultant acceleration is obtained using equation 3.1 (Bent, Cho, Wittmann, et al., 2021a).

$$a = \sqrt{a_x^2 + a_y^2 + a_z^2} \qquad 3.1$$

A summary of all the accelerometer-based biomarkers is presented in Table 3.6.

Accelerometer data can also indicate the type of activity being conducted. This requires heuristic methods or ML models to convert raw accelerometer data into activities of daily life. Features like activity bouts or activity estimates are key indicators of whether an individual is active or sedentary during the day, making them highly useful for predicting glucose-related outcomes.

Table 3.6: Activity related biomarkers

Feature	Definition	References
Accelerometer Statistics	Mean, Mode, 25th, 75th Percentile, range, maximum difference, minimum difference, daily averages, skewness and	(Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a;

	standard deviations of the accelerometer data.	Laguna Sanz et al., 2019; Zhu et al., 2022)
Activity Estimates	Based on some heuristics and combining HR and accelerometer data, some activity estimates are also used. Some studies also use the estimates from the proprietary software of the smart watches (Bertachi et al., 2020)	(Bent, Cho, Henriquez, et al., 2021)
Percentage of activity types	Using (*OxSmart watches/biobankAccelerometerAnalysis*, 2014/2024) accelerometer data is converted into activity bouts into sedentary, moderate, walking, sleep, and light.	(Lam et al., 2021a)

To analyse these biomarkers the guiding questions in Table 3.1 are answered in the Table 3.7

Table 3.7: DACIA information synthesis table for activity biomarkers

Phase	Guiding Questions	Answers
Data	Which smart watch sensor data is used to monitor relevant biomarkers? Which other data types can meaningfully complement the smart watch sensor data for higher digital biomarker accuracy? What could contribute to the missing data?	Activity biomarkers are measured using accelerometers. Empatica E4 , Fitbit are used in the studies identified. Activity estimates can be improved by combining accelerometer data with HR data (Fujimoto et al., 2013). (Lam et al., 2021a) augmented the accelerometer with Sex, Age, and other demographic information. Missing data can come from not wearing the sensing device
Aggregation	What is the time window length suitable for the biomarker estimate? Is the sampling frequency of sensor suitable for the desired window length?	Identified works have used 5-minute windows (Bent, Cho, Henriquez, et al., 2021) and 3 second (Jahromi et al., 2023) to 30 second (Lam et al., 2021a) windows with 50% overlap to calculate these biomarkers. (Jeon et al., 2019) used time lagged features for prediction of glucose. The sampling frequency of sensors in Empatica E4 is 32Hz and has been used for activity related features effectively
Contextualization	Are the primary signals influenced by the timing of measurements (such as weekday or season) or by specific participant characteristics (like gender, age, or body mass index (BMI))?	Accelerometers may follow circadian cycles and time of day, because of effect of day-night cycle on activity (Lam et al., 2021a). Waist circumference has been shown to effect these biomarkers (Länsitie et al., 2021)
Interpretation	What degree of change in signals related to the outcome of interest would be deemed significant and clinically relevant when assessing digital biomarkers from smart watches with glucose markers?	The feature importance analysis by Bent, Cho, Henriquez, et al. (2021) found that activity features contribute approximately 17% to predicting glucose markers. (Lam et al., 2021a) has shown that activity features are just as important as sociodemographic features. A combination of these features increases the performance of prediction models by 7-8%
Actions	What are considered meaningful interventions in response to significant deviations in digital biomarkers?	Significant changes in activity biomarkers is a strong predictor of glucose regulation. (Zahedani et al., 2023) shows that increased activity increases the prevalence of normo glycaemic values.

3.5.3.1 Practical steps for accelerometer data

To calculate the biomarkers from accelerometer data, the first step is filtering the data to overcome noise. Table 3.8 lists the filtering techniques used in the results of the literature review.

Table 3.8: Filtering techniques of preprocessing data

Filter	Details	Frequency Range
Butterworth low pass filter	Attenuates frequencies above a chosen cutoff frequency, removing high-frequency noise while preserving low-frequency signals.	Cut off: 30 Hz (Jahromi et al., 2023) Cut off: 20 Hz (Lam et al., 2021a)
Median filter	Applied to smooth data and remove outliers	(Sevil et al., 2021)
Savitzky-Golay filter	Fits a polynomial to each data window to preserve signal features	(Sevil et al., 2020)

The Savitzky-Golay filter enhances data quality by fitting a polynomial to each window of data points, allowing it to preserve the overall shape and key features of the signal. This is particularly important for capturing activity patterns that are directly linked to glucose level fluctuations. In contrast, the median filter is highly effective at removing outliers and sharp spikes by replacing each data point with the median of surrounding values. This process significantly reduces noise caused by sudden, erratic movements without distorting the underlying signal. When used together, these filters complement one another, improving the quality of accelerometer data and ultimately leading to more accurate glucose predictions. Accelerometer signals for human activity recognition lies within 0.1 to 20Hz (Twomey et al., 2018). A recent study suggests that hand tremors for hypoglycaemia detection fall within the range of (4-14 Hz) the so called hand tremor frequency range (HTFR) for T1DM patients (Abbas et al., 2018). Hence the cutoff frequency of 30 Hz is more suited.

If a public application programming interface (API) is used for calculating the biomarkers, accelerometer data should be converted to the relevant form for example, Empatica E4 records acceleration data in units of 1/64 Gs whereas other devices such as apple watch record these values in units of g. Hence relevant conversions must be done before calculating the biomarkers. Some of the studies use the acceleration components in x, y and z directions whereas other use resultant acceleration defined by equation 1 and the effect of gravitational acceleration G is removed from this resultant acceleration(Lam et al., 2021a).

The data is subsequently converted into windows of interest, they range from 3 seconds to 5 minutes in the literature with no or 50% overlap. This choice is dependent on the application and sampling frequency. When accelerometer data is to be converted into physical activity levels 3 second window with 50 % overlap is used. When the same data is to be used to detect hand tremors, a 30 second window without overlap is used. A 5-minute window is used when statistical features are used for CGM values prediction. A combination of these window sizes can be used for further feature engineering.

3.5.3.2 Activity Biomarker Importance in glucose prediction

The studies identified in the literature review highlight the importance of activity biomarkers in glucose prediction models. (Bent, Cho, Henriquez, et al., 2021) show that activity related biomarkers contribute 17% in prediction of hypoglycaemia glucose markers. As (Jahromi et al., 2023) uses hand tremors to identify hypoglycaemia events in T1DM patients, most features within HTFR range (0.4-14 Hz) are shown to be more significant based on impurity based feature importance (average band power :0.18 , maximum: 0.13, and minimum: 0.11). (Jeon et al., 2019) combines self-reported features such as meals, finger-stick glucose, illness, stress, exercise, and work of which two are activity features and found that these in combination with CGM features outperform other feature combinations. (Sevil et al., 2021) shows that addition of activity biomarkers reduced the mean absolute error from 35.1 to 31.9 mg/dL. (van Doorn, Foreman, et al., 2021) show that additional activity markers reduces the prediction error from 0.288 to 0.271 highlighting the importance of activity markers. (Laguna Sanz et al., 2019) shows that to enhance the accuracy of CGM values during movement the HR, metabolic estimates, and temperature values are more important than movement.

3.5.4 Physiological Biomarkers

Physiological biomarkers are significant predictors of glucose markers. The sensors on the smartwatches measure vital physiological data. Table 3.9 summarizes the smartwatches and the sensors that measure these physiological data as seen in the records identified in our literature review.

A standard photoplethysmography (PPG) sensor has a light emitter and detector. The emitter shines light onto the tissue, while the detector measures the amount of reflected light. The amount of reflected light is proportional to changes in blood volume. This helps measure Blood Volume Pulse (BVP), which tracks the periodic changes in blood volume associated with each heartbeat. HR is the number of heart beats per minute measured using the periodic peaks in the PPG signal.

The inter beat interval (IBI) is the time gap between individual heartbeats and is used to estimate instantaneous HR. The IBI sequence provided by Empatica is derived from the processing of the PPG/BVP signal, using an algorithm that removes incorrect peaks

caused by noise in the BVP signal. Additionally, EDA sensors measure the electrical conductance of the skin, which varies with sweat gland activity and provides information on emotional and stress responses.

Table 3.9: Sensors used to measure physiological biomarkers

Device	Sensors	References
Emaptica E4	Temperature, EDA, PPG as well as BVP inferred from PPG signal.	(B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a; Cho et al., 2023a; Zhu et al., 2022)
Zephyr Bioharness	Breathing rate.	(Alhaddad et al., 2022; Kumari et al., 2023)
Fitbit	HR, HRV and in some cases oxygen saturation, based on these features many sleep related data events data is available from Fitbit but have not been used in IG prediction pipelines	(Benedetti et al., 2021)
Garmin Vivoactive	HRV, HR, Motion, and time stamps.	(Lehmann et al., 2023)

EDA sensors are primarily of two types: Galvanic Skin Response (GSR) sensors and Skin Conductance Response (SCR) sensors. GSR sensors measure the overall electrical conductance of the skin, reflecting changes in sweat gland activity. SCR sensors, on the other hand, focus on transient changes in skin conductance in response to specific stimuli, providing insights into rapid emotional and physiological responses. Both types leverage the fact that skin conductance increases with sympathetic nervous system activation, offering valuable data on an individual's emotional and stress levels. Temperature sensors complement these measurements by monitoring body temperature.

The features are also calculated and used in the form of moving or rolling averages for the same reasons mentioned before. All these features can also be indicative of the ANS tone, but a separate section provides a more descriptive analysis of ANS biomarkers. HR data combined with accelerometer data can be used to heuristically determine the kind of activity (aerobic or anaerobic), which can be very useful in estimating the metabolic state of body, which can be a useful predictor of IG labels.

Based on these data streams various features of the physiological data are used in IG events prediction pipelines. Table 3.10 summarizes some of these.

Table 3.10: Physiology related features

Features	Definition	Sources
Temperature (statistical features)	Mean, Mode, 25th, 75th Percentile, range, maximum difference, minimum difference, daily averages, skewness and standard deviations of the temperature data.	(Bent, Cho, Henriquez, et al., 2021; Lehmann et al., 2023)
EDA (statistical features)	Mean, Mode, 25^{th}, 75^{th} Percentile, range, maximum difference, minimum difference, daily averages, skewness and standard deviations of the EDA data.	(Lehmann et al., 2023)
Breathing rate (statistical features)	Mean, Mode, 25^{th}, 75^{th} Percentile, range, maximum difference, minimum difference, daily averages, skewness and standard deviations of the breathing data.	(Kumari et al., 2023)
Heart Rate	HR from Fitbit has a different sampling frequency than that of Empatica hence HR change over time can be estimated using Fitbit but not the HR variability that are used in studies that use Empatica. Other statistical features are the same for HR as well.	(Site et al., 2023; van Doorn, Foreman, et al., 2021)

To synthesize the information for this category, the guiding questions of the DACIA framework are answered in the Table 3.11

Table 3.11: Answers to the DACIA guiding questions for Physiology Biomarkers

Phase	Guiding Questions	Answers
Data	Which smart watch sensor data is used to monitor relevant biomarkers? Which other data types can meaningfully complement the smart watch sensor data for higher digital biomarker accuracy? What could contribute to the missing data?	Physiological biomarkers are measured using PPG, breathing, EDA and temperature sensors. Empatica E4, Garmin Vivoactive, Fitbit and Zephyr Breathing devices are used in the studies identified. Physiological biomarkers are (Lam et al., 2021a) augmented the accelerometer with Sex, Age, and other demographic information. Missing data can come from not wearing the sensing device
Aggregation	What is the time window length suitable for the biomarker estimate? Is the sampling frequency of sensor suitable for the desired window length?	Identified works have used 5-minute windows (Bent, Cho, Henriquez, et al., 2021) and 3 second (Jahromi et al., 2023) to 60 second (Föll et al., 2021; Lehmann et al., 2023) windows. The sampling frequency of HR, BVP and IBI from Empatica E4 is 1Hz, 64 Hz and 1 Hz respectively whereas the temperature sensor is 4Hz.
Contextualization	Are the primary signals influenced by the timing of measurements (such as weekday or season) or by specific participant characteristics (like gender, age, or body mass index (BMI))?	Physiological biomarkers follow circadian cycles and time of day, because of effect of day-night cycle on activity (Lam et al., 2021a). Waist circumference has been shown to effect these biomarkers (Länsitie et al., 2021)
Interpretation	What degree of change in signals related to the outcome of interest would be deemed significant and clinically relevant when assessing digital biomarkers from smart watches with glucose markers?	The feature importance analysis by Bent, Cho, Henriquez, et al. (2021) found that activity features contribute approximately 13.4% to predicting glucose markers. Shapely additive explanations found the relative importance of HR related features is 32.8% in (Lehmann et al., 2023). The odds analysis of the features measured in (Zhu et al., 2022) show that IBI, and mean temperature have significant association with hypoglycaemia and hyperglycaemia.
Actions	What are considered meaningful interventions in response to significant deviations in digital biomarkers?	Significant changes in physiological biomarkers are a strong predictor of glucose regulation.

3.5.4.1 Preprocessing for physiological biomarkers

The preprocessing steps can be broken down into three basics steps: 1) Reading the data, 2) Noise Filtering and 3) Selection of Windows for biomarker calculation. For training the glucose marker prediction models, the time stamped data must be stored offline. Smart watches have the capacity to store the raw data from the sensors in flash memory for a finite amount of time depending on the device. For example, Empatica E4 can store 60 hours of data (McCarthy et al., 2016), Fitbit stores the data for a week (Feehan et al., 2018) whereas Garmin vivo active stores the data for a couple of weeks (Nikam & Mathew, 2020). Each of these devices have their respective platforms to access and save the data for offline usage. Recent works such as Closing the Loop on AI & Data Collection (CLAID) (Langer et al., 2024) and my personal health dashboard (MyPHD) enable data collection from multiple devices at once (Bahmani et al., 2021).

A comprehensive comparison of various such solutions is given in (Langer et al., 2024). The issues of interoperability, frequency differences and time asynchronization should be handled while storing the data for further usage.

Table 3.12: Preprocessing libraries for physiological biomarker preprocessing

Library	Reference	Language
biosppy	(Bota et al., 2024)	Python
HeartPy	(van Gent et al., 2019)	Python
hrv-analysis	(Niskanen et al., 2004)	Python
NeuroKit2	(Makowski et al., 2021)	Python
pyHRV	(Gomes, 2022)	Python
pyphysio	(Bizzego et al., 2019)	Python
pySiology	(Gabrieli et al., 2020)	Python
FLIRT	(Föll et al., 2021)	Python
Smart watches	(de Looff et al., 2022)	R
Bio-SP	(Nabian et al., 2018)	Matlab

Publicly available APIs can carry out the noise filtering and windowing steps. Some of these APIs referred to in the works identified in this literature review are given in Table 3.12. These libraries handle different aspects of the preprocessing and biomarker calculations for the various smart watches. The preprocessing steps are categorized into outlier removal and noise filtering. In outlier removal infeasible sensor values are detected based on their mean, standard deviation and relative change in successive values. The sources of noise depend on the method of measurement in sensing devices. HR measurements based on PPG sensors are shown to be robust to skin tone, but the type of smart watches and motion cause noise in HR measurements (Bent et al., 2020).

The sources of noise in HR measurements are suspected to be: motion artefacts, and environmental factors that result in missing or false beats (Castaneda et al., 2018). Especially cyclical motion causes a signal cross over resulting in misclassification of periodic signal from movement as cardiovascular cycle(Reis et al., 2019).

To filter out noise, literature suggests several methods: bandpass filtering within the range of 0.5 Hz to 4 Hz to cover normal physiological HR while excluding low-frequency components and high-frequency noise ; moving average filters to smooth short-term fluctuations ; wavelet-based denoising to separate noise based on frequency and temporal location ; adaptive filtering for non-stationary noise sources ; and Kalman filtering (B. S. Kim & Yoo, 2006; Q. Li et al., 2008; Temko, 2017) . IBI values are usually measured using BVP and HR values and carry over their noise.

IBI values have the same noise sources as HR values as both are measured using PPG sensors but they manifest in the form of ectopic beats or heart beats missed by the sensor (Jovanov, 2015). For outlier removal in IBI values successive values and their relative differences used in (Föll et al., 2021). For temperature sensors common sources of noise include environmental conditions such as changes in ambient temperature, which can cause fluctuations in sensor readings. This is particularly relevant for smart watch devices used in both indoor and outdoor settings, where temperature changes can be significant due to factors like weather and air conditioning. Other sources of noise include physiological variations, such as perspiration. Additionally, motion artifacts caused by physical activity can introduce noise (Ellebrecht et al., 2022). Temperature sensors have different sources of noise depending on the kind of sensor, for example Empatica E4 and Fitbit has Infrared Thermopile sensors. There are studies that examine the noise sources in ear thermopile sensing devices (Mbarek et al., 2022). There exists a need to systematically identify the noise sources and their mitigation strategies for such sensors. Some relevant studies have suggested using Hampel filter and Savitzky-Golay filter (Aryal & Becerik-Gerber, 2019).

BVP values are filtered in the literature with 4^{th} order Butterworth band pass filter with allowed range 1Hz to 8 Hz but that is not applied in glucose prediction studies (Y. Xu et al., 2017). Table 3.13 lists preprocessing steps carried out for different sensors.

3.5.4.2 Biomarker importance in glucose prediction models

Physiological biomarkers are effective predictors of glucose values. (Zahedani et al., 2023) performed the ablation analysis to find that after nutrient information HR data has the most significant impact on CGM value prediction.

(Bent, Cho, Henriquez, et al., 2021) show that physiological features contribute 13.6% to CGM value prediction (Temperature = 5%, HR = 3.2%, EDA =5.4%). For HbA1C

prediction (Bent, Cho, Wittmann, et al., 2021a) found the relative importance of skin temperature is 33%, EDA is 28%, and HR is 14%.

Table 3.13 preprocessing techniques for different sensors

Sensor	Preprocessing	Libraries
HR	Band pass filter [0.5,4Hz], Moving Average filter, Wavelet filter, Kalman Filtering.	--
IBI	Outlier Removal, Successive IBI values should be within [20%, 25.5%,32.5%].	FLIRT, Bio-SP
EDA	Outlier Removal, Noise Filtering	FLIRT, Bio-SP
Temperature	Hampel Filter, Savitzky-Golay filter	--
BVP	Bandpass Filter [1Hz-8Hz]	

3.5.5 Autonomic Nervous System Biomarkers

In the records that have come out of the literature search the most common thread is the use of ANS related biomarkers. ANS is divided into sympathetic (SNS) and parasympathetic system (PNS). SNS is a fight and fight response characterized by elevated HR. The blood glucose level increases during sympathetic arousal (Lundqvist et al., 2019). Whereas PNS controls the rest and digest functions (Debnath et al., 2021). ANS biomarkers are drawn from the HR, EDA and temperature sensors or a combination of those. Table 3.14 lists the ANS biomarkers used for glucose prediction in the recorded works.

The information about these biomarkers is synthesized using the DACIA framework in the Table 3.15.

Table 3.14: Automatic Nervous System related features

Features	Definition	References
Heart Rate Variability (HRV) (Statistics)	HRV, Mean, Mode, 25th, 75th Percentile, range, maximum difference, minimum difference, daily averages, skewness and standard deviations of the HRV data.	(B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a; Kumari et al., 2023)
Inter beat Interval	IBIis the time between two successive heart beats it is also called NN interval.	(B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a; Kumari et al., 2023),(Alhaddad et al., 2022)
Standard deviation of NN (SDNN)	The standard deviation of NN interval called SDNN is also used as a feature in glucose level predictors, it also measures the PNS activity.	(B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a; Kumari et al., 2023)
Root mean square of successive differences (NN) (RMSSD)	The root mean square of the successive NN intervals is also used. It measures the combined effect of vagal and sympathetic ANS.	(B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a; Kumari et al., 2023)
pNN50	pNN50 is the percentage of intervals that differ by more than 50 milliseconds and is a metric to evaluate vagal activity.	(B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a; Kumari et al., 2023)
Peak EDA	Peak value of EDA is measured by setting the distance between peaks is required to be a second with a prominence of 0.3 micro siemens. Various statistical derivatives of peak EDA are used in various work. Peak EDA measures the SNS activation.	(B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021a; Kumari et al., 2023)

Table 3.15 DACIA questions for ANS biomarkers

Phase	Guiding Questions	Answers
Data	Which smart watch sensor data is used to monitor relevant biomarkers? Which other data types can meaningfully complement the smart watch sensor data for higher digital biomarker accuracy? What could contribute to the missing data?	ANS biomarkers are measured using PPG, breathing, and EDA sensors. Empatica E4, Fitbit is used for ANS biomarkers in the studies identified. ANS biomarkers are augmented the accelerometer values (M. Wu et al., 2015), with Sex (Voss et al., 2015), Age (J. Zhang, 2007), Temperature (Yoo & Chung, 2018) and other demographic information. Missing data can come from not wearing the sensing device.
Aggregation	What is the time window length suitable for the biomarker estimate? Is the sampling frequency of sensor suitable for the desired window length?	Identified works have used 5-minute windows (Bent, Cho, Henriquez, et al., 2021) and 3 second (Jahromi et al., 2023) to 60 second (Föll et al., 2021; Lehmann et al., 2023) windows. The sampling frequency of HR, BVP and IBI from Empatica E4 is 1Hz, 64 Hz and 1HZ whereas the temperature sensor is 4Hz.
Contextualization	Are the primary signals influenced by the timing of measurements (such as weekday or season) or by specific participant characteristics (like gender, age, or body mass index (BMI))?	ANS biomarkers follow circadian cycles and time of day (Boudreau et al., 2013), because of sleep. Weight has been shown to effect these biomarkers (Alrefaie, 2014)
Interpretation	What degree of change in signals related to the outcome of interest would be deemed significant and clinically relevant when assessing digital biomarkers from smart watches with glucose markers?	The feature importance analysis by (Bent, Cho, Henriquez, et al., 2021) found that ANS features contribute approximately 8.3% to predicting glucose markers. Shapely additive explanations found the relative importance of ANS features is 27.3% in (Lehmann et al., 2023).
Actions	What are considered meaningful interventions in response to significant deviations in digital biomarkers?	ANS biomarkers interact with food related features as well as circadian biomarkers in glucose prediction studies.

3.5.6 Circadian Cycle biomarkers

The circadian rhythm of the body is also used in a couple of studies (Bent, Cho, Henriquez, et al., 2021; Bent, Cho, Wittmann, et al., 2021) using hours and minutes from midnight as a feature to be used in ML models for IG prediction. They have proven to be one of the most important features in those studies. Other studies have explored the ultradian cycles in the context of IG levels (Shannahoff-Khalsa, 2007). IG management is not just affected by the ultradian cycles of the IG levels, they are also influenced by other cyclical process in the body (Crofts et al., 2018). (Zahedani et al., 2023) took the sine and cosine of the clock to ensure temporal continuity.

Table 3.16: Answers to the DACIA guiding questions for Circadian Biomarkers

Phase	Guiding Questions	Answers
Data	• Which smart watch sensor data is used to monitor relevant biomarkers? • Which other data types can meaningfully complement the smart watch sensor data for higher digital biomarker accuracy? • What could contribute to the missing data?	• Circadian biomarkers are measured with the time stamps of the data. • Circadian biomarkers are not dependent on other data types but features can be engineered that can capture different ultradian cycles within the data. • Missing data can come from not wearing the sensing device.
Aggregation	• What is the time window length suitable for the biomarker estimate? • Is the sampling frequency of sensor suitable for the desired window length?	• The window size is the same as the prediction window. • Circadian biomarkers are independent of the sampling frequency.
Contextualization	• Are the primary signals influenced by the timing of measurements (such as weekday or season) or by specific participant characteristics (like gender, age, or body mass index (BMI))?	• Circadian biomarkers are mostly independent of external features
Interpretation	• What degree of change in signals related to the outcome of interest would be deemed significant and clinically relevant when assessing digital biomarkers from smart watches with glucose markers?	• The feature importance analysis by (Bent, Cho, Henriquez, et al., 2021) found that circadian biomarkers contribute approximately 10.6% to predicting glucose markers. Ablation analysis of features by (Zahedani et al., 2023) also highlights the importance of circadian features.
Actions	• What are considered meaningful interventions in response to significant deviations in digital biomarkers?	• Circadian biomarkers interact with other biomarkers such as food and HR in glucose prediction models.

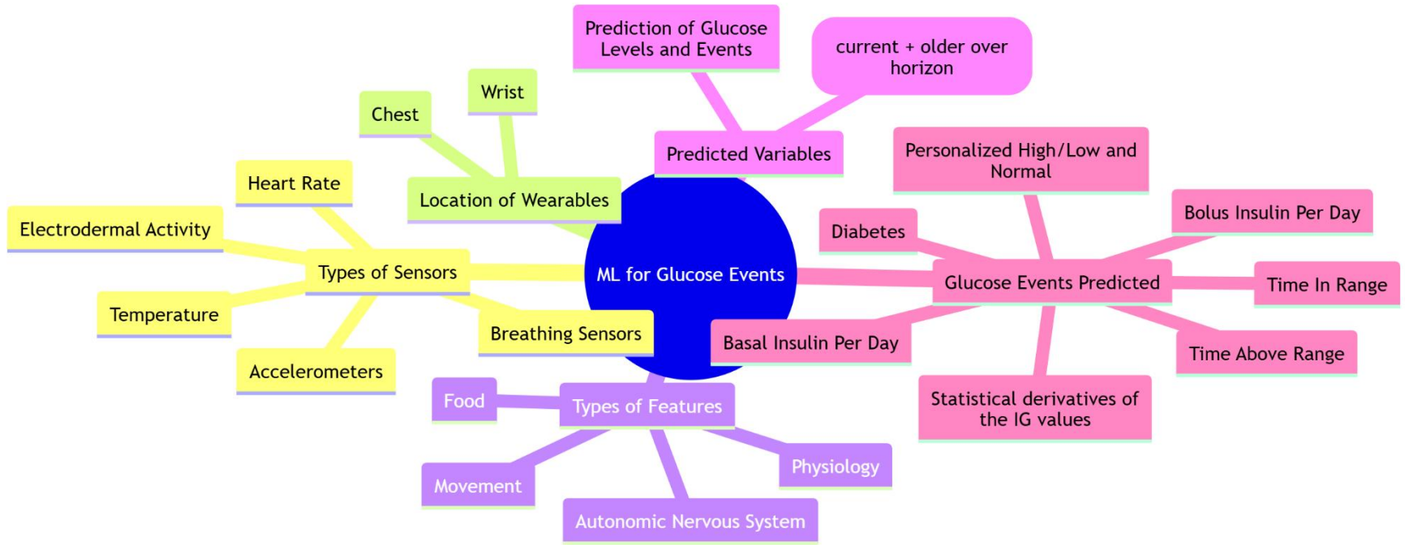


Figure 3.7: Summary of various aspects of the analysed studies.

Figure 3.7 summarizes the various aspects of the studies that have been analysed in this work.

3.6 Identified Gaps

While sleep related metrics (e.g. sleeping time) can be drawn from sensors such as accelerometers as well as a combination of accelerometers and HR sensors, these models do not use the sleep features, hence an investigation into the utility of sleep features for IG prediction must be tested.

The ANS are used in these works, however these are not the only ANS markers that can be identified using this data and a need thus arises to compare the effectiveness of other such features, that may include various versions of pNN features (pNN10,pNN30,pNN50, and pNN70). A Fourier transform of the HRV signal, and calculating the power spectral density of 0.15-0.4 Hz components is called high frequency power (HF) and it measures parasympathetic activity, and that between 0.04 to 0.15 Hz is considered to LF power. HF is related to PNS whereas LF is related to ANS. A ratio of these two measures symathovegal balance that can be used as a feature of individuality. For data that includes the breathing rate (Marling & Bunescu, 2020), respiratory sinus arrythmia (RSA), which measures the synchronous respiratory cycle length and IBI, for the same purpose. Powers from very low frequency can also be used to incorporate breathing.

While most models are ML models in this chapter, there are other DL models that can be viewed to test the kind of layers a foundational DL model from the different data streams may be constructed.

Finally, more work can be done in the domain of representation learning so that less data hungry models can be developed.

While circadian features are used in these works, a need also arises to compare the effect of ultradian cycles on the IG.

3.7 Conclusion

In this Chapter, firstly, the relevant records are identified by using a well-crafted search term and then searching for the records on scientific databases such as PubMed and Scopus to identify the relevant records. These records are then subjected to a Prisma review process. The relevant records are identified based on inclusion and exclusion criteria.

After analysing the identified records, we aimed to find the biomarkers and labels that are used in these works. We identified that the values predicted are not just limited to IG levels and values but also their statistical derivatives as well as clinically relevant values (e.g TIR, Time Above Range). These values encompass various aspects of the health of

the individual. For example, personalized high, low and average is used in prediabetics and healthy populations whereas events such as hyper and hypoglycaemia are predicted for people with metabolic disorders.

The features used in these works are varied and fall under 5 broad categories. These categories include food, physiology, circadian rhythm, movement, and ANS. While there is only one study that specifically used the circadian rhythm in the form of a running clock as an input feature to the model, they found that for the DT they trained it was the most important feature. Food and drink affect the IG levels and are also very important features, thus any work that would find an easier mechanism to log the food values effectively or predict the food values based on other sensors would be very valuable. Features that measure ANS activity are also found to be very useful. Since most other features can be drawn from physiological features, those features are also found to be used in a variety of these works.

The work then goes on to identify the gaps and finds that the sleep features have not been used as extensively as other features even though sleep has a reported impact on the IG levels.

Finally, we conclude this Chapter with a discussion on the results.

Table 3.17 Summary of the related articles

Titles	Reference	Population	Datasets	Models	Validation Data	Labels	Evaluation Metrics
Advanced Diabetes Management Using Artificial Intelligence and Continuous Glucose Monitoring Sensors	(Vettoretti et al., 2020)	Type 1 Diabetes and predicts bolus insulin intake OhioT1DM	(In Silico) Padova OhioT1DM	NN Regularized Multiple Linear Regression Autoregressive Models with Exogenous input (ARX),	20%	Parameters for exogenous bolus glucose, blood glucose, post prandial hypoglycaemia , nocturnal hypoglycaemia & Glucose concentrations, Food log and Insulin Pump	Root mean square error (RMSE) = (19-20 mg/dL)
Hypoglycaemia Detection Using Hand Tremors: Home Study of Patients With Type 1 Diabetes	(Jahromi et al., 2023)	33 Type 1 Diabetes Mellitus (T1DM) patients	Private Data	SVM K Nearest Neighbours (KNN) Random Forest (RF)	Leave one out of cross validation (LOOCV)	hypoglycaemia	Precision (KNN :82.03%, SVM :81.48%, RF: 80.37%) Recall (KNN:82.03%, SVM:81.48%, RF: 80.37%) F1-Score: (KNN:79.41%, SVM:78.28%, RF: 78.88%)
In-human testing of a non-invasive continuous low-energy microwave glucose sensor with advanced machine learning capabilities	(Kazemi et al., 2023)	Takes in look back values and outputs look ahead values.	4 healthy and 2 diabetic patients	LSTM	20%	CGM values and daily variation	Mean Squared Error: 0.0849 of amplitude [dB]
Non-invasive smart watches for remote monitoring of HbA1c and glucose variability: proof of concept	(Bent, Cho, Wittmann, et al., 2021a)	Predicts levels of glucose (HbA1C) for the first day as well as glucose variation	26 Healthy Participants 16 for training and 10 for validation	27 Random Forest Models were developed to predict various Glucose markers.	Leave one out of cross validation (LOOCV)	CGM values and daily variations	Best (glucose management indicator; RMSE=0.26 mg/dL), Worst (mean of normal glucose RMSE= 10.15) &

			(Healthy and Prediabetics)				CGM values and daily variations
Towards Smart watch-based hypoglycaemia Detection and Warning in Diabetes	(Maritsch et al., 2020)	Predicts hyperglycaemia at individual level	One healthy and One participant with T1DM	Gradient Boosted Decision Tree (GBDT)	Stratified 10-fold cross-validation	hypoglycaemia	Mean accuracy of 82.7%
Early Detection of hypoglycaemia in Type 1 Diabetes Using Heart Rate Variability Measured by a Smart watch Device	(Olde Bekkink et al., 2019)	Detects hypoglycaemia using HRV	23 participants with T1DM	Forward step wise linear model	Statistical Measures	hypoglycaemia	Accuracy = 82%
Noninvasive hypoglycaemia Detection in People With Diabetes Using Smartwatch Data	(Lehmann et al., 2023)	Detects hypoglycaemia using HRV	31 Participants with T1DM	GBDT	30%	hypoglycaemia	AUROC = 0.76 ± 0.07
Enhanced Accuracy of Continuous Glucose Monitoring during Exercise through Physical Activity Tracking Integration	(Laguna Sanz et al., 2019)	Error between plasma glucose value and CGM value for activities was predicted.	6 Patients with Type 1 Diabetes	Linear Regression Models to predict the difference between CGM and plasma glucose values.	20%	CGM Error	Mean Absolute Relative Deviation 13.6%, Clarke Error Gird Analysis.
Prediction of Nocturnal hypoglycaemia in Adults with Type 1 Diabetes under Multiple Daily Injections Using Continuous Glucose Monitoring and Physical Activity Monitor	(Bertachi et al., 2020)	Predicting nocturnal hypoglycaemia events.	T1 Diabetics.	Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP).	K Fold Cross Validation K=5	Nocturnal hypoglycaemia	best (SVM Sensitivity=78.75%), worst (SVM Sensitivity=61.3%)
Utility of Big Data in Predicting Short-Term Blood Glucose Levels in Type 1 Diabetes Mellitus Through	(Rodríguez-Rodríguez, Chatzigiannakis, et al., 2019)	Look behind values were used to predict look ahead values	T1 Diabetics	Autoregressive integrated moving average (ARIMA), Random Forest (RF), SVM	(15 min, 30 min, 45 min and 60 min windows)	CGM values	(15 min) RMSE< 11.65mg/dL (RF) and RMSE = 22mg/dL

Machine Learning Techniques							
Physical Activity and Psychological Stress Detection and Assessment of Their Effects on Glucose Concentration Predictions in Diabetes Management	(Sevil et al., 2021)	CGM Forecasting	12 participants with T1DM	Recursive version of the prediction-based subspace identification (rPBSID) algorithm	1,1.5, 2 and 4 hours	CGM values	Mean Absolute Error (1H: 7.5 mg/dL, 1.5H: 10.7mg/dL, 2H: 37.5 mg/dL, 4H: 125 mg/dL)
On the Possibility of Predicting Glycaemia 'On the Fly' with Constrained IoT Devices in Type 1 Diabetes Mellitus Patients	(Rodríguez-Rodríguez, Rodríguez, et al., 2019)	Implemented the aforementioned models in (Rodríguez-Rodríguez, Chatzigiannakis, et al., 2019) on Edge devices	T1 Diabetes	ARIMA, SVM and SF	15 min, 30 min, 45 min and 60 min windows)	CGM values	Root Mean Squared Error (RMSE) = 11.65 mg/dL with 15 minutes horizon,
Classification of Postprandial glycaemic Status with Application to Insulin Dosing in Type 1 Diabetes-An In Silico Proof-of-Concept	(Cappon et al., 2019)	High, Low, Normal Glucose levels prediction	T1DM (Padova)	Extreme Gradient Boosting (XGBOOST)	K Fold Cross Validation K=3	CGM values	Area under receiver operator curve (AUROC) = 0.84
Machine-Learning-Based Diabetes Prediction Using Multisensory Data	(Site et al., 2023)	Prediction of Diabetes	D1Namo Dataset (Dubosson et al., 2018)	XGBOOST	20%	Diabetes	Accuracy = 98.2%
A computational framework for discovering digital biomarkers of glycaemic control	(Bartolome & Prioleau, 2022)	Predicted 10 markers of glycaemic control	250 subjects with type 1 and 2 Diabetes (*Big Data Donation Project* | *Tidepool*, n.d.)	SVM, DT and Linear Discriminant Analysis (LDA) Classifier	20%	Glycaemic Control	F1 Score= 89.7%
Machine learning-based glucose	(van Doorn, Foreman, et al., 2021)	Prediction of glucose values	Private data and translated the	ARIMA, Support Vector Regression (SVR), Light	20%	CGM values	RMSE= 0.19mmol/L with 15-minute horizon and

prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study		over 15 minutes and 60 minutes horizon	results onto OhioT1DM (Marling & Bunescu, 2020)	gradient boosting, shallow MLP, RNN and LSTM			0.59 mmol/L with 60 minutes horizon
Improving the Accuracy of Continuous Blood Glucose Measurement Using Personalized Calibration and Machine Learning	(Kumari et al., 2023)	Predicting CGM values	D1Namo (Dubosson et al., 2018)	SVR, KNN, DT, RF, AdaBoost and MLP	20%	CGM values	MARD = 8.3%

3.9 Chapter Summary

This chapter presents a comprehensive examination of digital biomarkers measured using smart watch for their potential integration into ML models of glucose markers prediction. Through this review, different biomarkers and the related preprocessing steps were identified. In Chapter 2 a SLR of the broader research landscape was conducted, which gives a technical comparison of different methods used in developing ML pipelines from problem description to evaluation of the model structured under various paradigms effectively answering research question one. Chapter 3 is a more targeted treatment of various digital biomarkers that are used for prediction of IG developed using smart watch and food log data. This answers the research question two of this thesis. It identifies a series of digital biomarkers that are used in similar studies and identifies two main gaps; different ML models of IG prediction need to be compared based on empirical data, and sleep biomarkers that have reported relationship with IG values need to be investigated as input to ML models.

4 Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log

4.1 Preface

The content of this Chapter is based on the Journal article *"**Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log**"*, published and peer-reviewed in the Journal Electronics Online. All the sections of the article were expanded to provide a more extensive content and discussion on the effectiveness of different machine learning (ML) models in predicting interstitial glucose (IG) levels from smart watch sensors.

Since the data from smart watches is measured longitudinally, it is prone to noise as data as participants can take off the watch or the relative motion between watch and skin. Given this noisy data from smart watch devices, it is important to recognize models that are robust to noise and have good performance (lower root mean square error (RMSE) and mean absolute error (MAE)) when predicting IG levels. This work systematically investigates which models have better performance in IG prediction for smartwatch data and food logs using difference between predicted and actual IG values determined by performance metrics like RMSE and MAE etc. It is also important to recognize that of the 69 features measured from smart watches identified in Chapter 3, which features are more informative to the models, and this work identifies the important features by explaining the results of the models in terms of the input feature ranges using SHAP values. It also highlights the complex interactions between features using partial dependence plots (PDPs) picked by tree models, highlighting them as baseline models for IG prediction tasks.

This work and its related experiments were significantly informed by results and findings from the extensive literature review of time domain methods in Chapter 2, the narrow review in Chapter 3. The outcome of Chapter 3 implies that ML models used for predicting IG levels need to be examined empirically using data. This Chapter uses empirical data and features (or biomarkers) and compares the models based on performance metrics.

4.2 Abstract

This Chapter examines the performance of various machine learning (ML) models in predicting Interstitial Glucose (IG) levels using data from wrist-worn smart watch sensors. The insights from these predictions can aid in understanding metabolic syndromes and disease states. A public dataset comprising information from the Empatica E4 smart

watch, the Dexcom Continuous Glucose Monitor (CGM) measuring IG, and a food log was utilized. The raw data were processed into features, which were then used to train different ML models. This study evaluates the performance of decision tree (DT), support vector machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA), K-Nearest Neighbours (KNN), Gaussian Naïve Bayes (GNB), lasso cross-validation (LassoCV), Ridge, Elastic Net, and XGBoost models. For classification, IG labels were categorized into high, standard, and low, and the performance of the ML models was assessed using accuracy (40–78%), precision (41–78%), recall (39–77%), F1-score (0.31–0.77), and receiver operating characteristic (ROC) curves. Regression models predicting IG values were evaluated based on R-squared values (−7.84–0.84), mean absolute error (5.54–60.84 mg/dL), root mean square error (9.04–68.07 mg/dL), and visual methods like residual and QQ plots. To assess whether the differences between models were statistically significant, the Friedman test was carried out and was interpreted using the Nemenyi post hoc test. Tree-based models, particularly RF and DT, demonstrated superior accuracy for classification tasks in comparison to other models. For regression, the RF model achieved the lowest RMSE of 9.04 mg/dL with an R-squared value of 0.84, while the GNB model performed the worst, with an RMSE of 68.07 mg/dL. A SHAP analysis identified time from midnight as the most significant predictor. Partial dependence plots revealed complex feature interactions in the RF model, contrasting with the simpler interactions captured by LDA.

4.3 Introduction

Diabetes is characterized by increased glucose levels. The incidence of diabetes is increasing at a rapid rate. According to the World Health Organization, the number of people with diabetes worldwide rose from 108 million in 1980 to 422 million in 2014 (Bent, Cho, Wittmann, et al., 2021a), and this number is projected to reach 700 million by 2045 (Home et al., 2021). Prediabetes is a series of risk factors of diabetes defined using fasting glucose levels between 100 and 125 mg/dL (Aguilar et al., 2015). Prediabetes affects approximately 34% of adults in the United States (Aguilar et al., 2015), with nearly 7.3 million undiagnosed cases (CDC, 2024b),(Grundy et al., 2004). However, 85% of individuals with prediabetes are unaware (Ervin, 2009) that they have it (Ford et al., 2008). Early intervention through lifestyle changes or medication can significantly reduce the risk of progression from prediabetes to diabetes by up to 58% (CDC, 2024a). Monitoring glucose levels is thus helpful for managing and preventing metabolic diseases (Jarvis et al., 2023). Classically, glucose levels are measured using a blood test that measures glycated haemoglobin levels (HbA1C). Fasting HbA1C levels measure glucose regulation for the past two to three months and do not measure fluctuations and short-term glucose spikes (Zoungas et al., 2012). Monitoring short-term glucose

variations is essential for adjusting medication, dietary habits, and physical activity to maintain optimal glucose regulation. To measure these short-term glucose spikes, continuous glucose markers (CGMs) are used. Glucose regulation markers such as time in range (TIR) can be measured using CGMs. CGMs are attached to the body with the help of a thread that penetrates the interstitial fluid. CGMs log Interstitial Glucose (IG) values every one to five minutes depending on the device. IG values are stored in them for up to 8 h. The stored IG values from CGMs can be downloaded with the help of Bluetooth technology (Beck et al., 2019). CGMs require regular downloading of the data and are minimally invasive.

In comparison to CGMs, smart watches are non-invasive and self-updating. Therefore, there is an increased interest in using smart watches for predicting IG levels. An example of this growing interest is the curation of various datasets (Cho et al., 2023b) that include smart watch data paired with glucose labels (Ali et al., 2023). Smart watches are equipped with sensors capable of tracking various physiological parameters. Smart watch sensors include heart rate, an accelerometer, and skin conductance, etc. The smart watch sensor values can be used to engineer predictors of IG (Adams & Nsugbe, 2021).

In addition to enhancing individual glucose management, predicting IG levels from smart watch sensors can also contribute to population-level health insights and disease management strategies. Aggregated data from smart watches and CGMs can provide valuable epidemiological information about glucose trends, the prevalence of metabolic conditions, and the impact of lifestyle factors on glycaemic control.

Machine learning (ML) algorithms are used to predict IG markers from smart watches due to their ability to extract complex patterns and relationships (Ali et al., 2024). ML models can adapt and improve over time by continuously learning from new data, making them well suited for personalized glucose monitoring and management (Zahedani et al., 2023). Studies have demonstrated the effectiveness of ML algorithms in predicting glucose levels from smart watch sensor data, achieving high levels of accuracy and precision (Ali et al., 2023, 2024; B. Bent et al., 2021; Bent, Cho, Wittmann, et al., 2021a). Many of these models add food log data, in addition to smart watch data.

In this study, we categorize IG values into high, low, and normal labels as described in (Ali et al., 2023; B. Bent et al., 2021; Bent, Cho, Henriquez, et al., 2021). Unlike traditional classifications of hyper- and hypoglycaemia, which are tailored for diabetic patients, this approach considers individualized glucose fluctuations. These designations are dynamic and personalized, reflecting an individual’s unique glycaemic baseline and accounting for circadian and intra-/inter-day variability.

4.3.1 Related Work

Earlier works utilized support vector machines (SVMs) and decision trees for predicting IG values and categories using smart watch data. For example, ref. (Bent, Cho, Henriquez, et al., 2021) produced 69 features predictive of glucose, defined the classification problem, and used DT to perform a classification with a root mean squared error (RMSE) equal to 21.22 ± 4.14 mg/dL. Similar works used depth vision guiding for recognizing human activity that can be used as input to glucose-monitoring models (Qi et al., 2022; J. Zhao et al., 2024). However, this requires additional sensors. Another work (Ali et al., 2023) recently designed four additional features using smart watch data but only performed a classification of CGM values into normal, high, and low and found support vector machine (SVM) to have an accuracy of 69% and DT to be 72.38% accurate. Another work utilizes extreme gradient boosting (XGBoost) models to classify the IG values of each participant with minimum accuracy = 60% and maximum accuracy = 86% (Ali et al., 2024).

While these works used smart watches and CGM data to train ML models to predict IG markers (classes and values), it would be useful to compare different ML models for both the classification and regression problem using the same performance metrics. While these works report a hyperparameter tuning process for the models, there is a need for hyperparameter tuning for the best performing models. To compare the performance of the ML models, model explanations and visual techniques can inform why certain models such as tree models outperform other models (Bent, Cho, Henriquez, et al., 2021). The comparisons of earlier works is given in Table 4.1.

Table 4.1: The performance of different models in related works.

Study	Type	Models	Performance
(Bent, Cho, Henriquez, et al., 2021)	Classification/Regression	DT	MSE = 21.22 ± 4.14 mg/dL
(Ali et al., 2023)	Classification	DT/SVM	Accuracy SVM (69%), DT (72.38%)
(Ali et al., 2024)	Classification	XGBoost	Accuracy (60–86%) *
(Maged & Atia, 2022)	Regression	Gradient Boosting	MSE = 23.40 mg/dL
(Lehmann et al., 2023)	Classification	DT	AUROC = 0.76 ± 0.07

(Huang et al., 2023)	Regression	RF	RMSE = 26.83 mg/dL
(Adams & Nsugbe, 2021)	Classification	SVM	Accuracy = 72.6 ± 2.4%

In summary, this work makes these novel contributions in comparison to other works:

- C1: Compare performance of ML models in predicting IG values using smart watch data as input and compared using Friedman test with Neymani post hoc analysis.
- C2: Compare performance of ML models in classifying IG values into high, low, and normal classes using smart watch data as input and compared using Friedman test with Neymani post hoc analysis.
- C3: Explain why different model types, such as tree-based models (RF, DT, and XGBoost), outperform SVM and GNB using partial dependence plots and Cook's plots.

In this context, our study compares several ML models, including DT, SVM, RF, Linear Discriminant Analysis (LDA), K-Nearest Neighbours (KNN), Gaussian Naïve Bayes (GNB), lasso cross-validation (LassoCV), Ridge, Elastic Net, and XGBoost. For the classification task, accuracy, precision, recall, F1-score, and ROC are used to compare models. Additionally, regression models predicting IG values are evaluated using R-squared values, mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and mean squared logarithmic error (MSLE). The hyperparameters of the best performing models are optimized using Bayesian Optimization with Optuna (Akiba et al., 2019). This work also explains why different models perform better than others, using partial dependence plots (PDP) to show feature interactions. This work also shows the robustness of RF models to influential outliers and the existence of such outliers using a Cook's plot.

The paper is organized as follows: Section 4.4 Materials and Methods**:** A detailed description of the dataset, highlighting its key characteristics and pertinent attributes, and an explanation of the data preprocessing, feature extraction techniques, and ML models used in this study. Section 4.5 Results: Presentation and comparison of the outcomes from the regression and classification analyses and their limitations. Sections 4.6 and 4.7 Discussion and Conclusion: Synthesis of the findings, discussion of their implications.

4.4 Materials and Methods

In this study, we utilized a public dataset (Cho et al., 2023). These data comprise a cohort of participants aged 35–65 years, inclusive, with elevated blood glucose levels falling

within the range of normal to prediabetic. The dataset consists of 9 female participants and 7 male participants. In this dataset, participants were required to wear a Dexcom G6 CGM and an Empatica E4 wristband for a duration of 8–10 days, during which physiological measurements such as heart rate (HR), inter beat interval (IBI), blood volume pulse (BVP), electrodermal activity (EDA), skin temperature, and tri-axial accelerometery were recorded.

Participants were also provided standardized breakfast meals every other day, and a food log was maintained. Date shifting was performed on the collected data to ensure participant de-identification.

The dataset includes a total of 16 participants. The files contain timestamped data. The ACC file provides accelerometer data for the X, Y, and Z orientations, while the BVP file records BVP measurements. The food log file documents the food items consumed by each participant, including details such as date, time, logged food, amount, calories, total carbohydrates, dietary fibre, sugar, protein, and total fat content. Demographics, including gender and HbA1C values for each participant, are also provided in the dataset given the supplementary material Table S1. The PPG is sampled at 64 Hz, giving the HR and BVP every second for IBI computation. The EDA and skin temperature are sampled at 4 Hz, and the accelerometery at 32 Hz. CGM records a value of IG every 5 min.

For preprocessing, the HR and IBI data were filtered with a Chebyshev II order-4 filter with a stopband attenuation of 20 dB and a passband of 0.5–5 Hz, as described in (Liang et al., 2018). For the removal of noise, a Gaussian low-pass filter was used with a sigma value of 400 ms (Nabian et al., 2018). We then segmented the filtered data into 5 min windows and aligned them with the Dexcom sensor values. Features were extracted from each window as described in (Bent, Cho, Henriquez, et al., 2021). These features can be broadly categorized into circadian features, statistical features of the sensor values, EDA features, and food features. A five-minute window was used for the calculation of these features, as the IG ground truth is available every five minutes. For the classification of IG labels, daily averages and standard deviations of CGM values were calculated. CGM values that are higher than the mean + standard deviation are considered high; conversely, values smaller than mean–standard deviation are considered low, whereas all the other values are considered normal, as described in (Bent, Cho, Henriquez, et al., 2021).

Accelerometer data were pre-processed using a Butterworth low-pass filter with cutoff frequency = 20 Hz, as explained in (Lam et al., 2021b). The resultant acceleration was calculated from the X, Y, and Z components and corrected for the gravitational

acceleration in the y direction. The EDA sensor data were smoothed to remove any artefacts. To do this, a Gaussian low-pass filter is used, with a 40-point window and value of sigma = 400 ms (Nabian et al., 2018). The HR data are filtered using a band-pass filter, filtering activity outside the [0.5–4 Hz] range. IBI data are filtered using a filter defined in (Föll et al., 2021). BVP data are filtered using a moving average smoothing filter, whereas temperature sensor data are filtered using a Savitzky-Golay filter (Chandra et al., 2021). This is illustrated in Figure 4.1.

The features are calculated in this work as described in (Ali et al., 2023; Bent, Cho, Henriquez, et al., 2021). The features are broadly categorized into four main classes: food features, circadian features, statistical features, and autonomic nervous system features. These features are calculated for each 5 min window. The mathematical definition of the features is provided in Table 4.2.

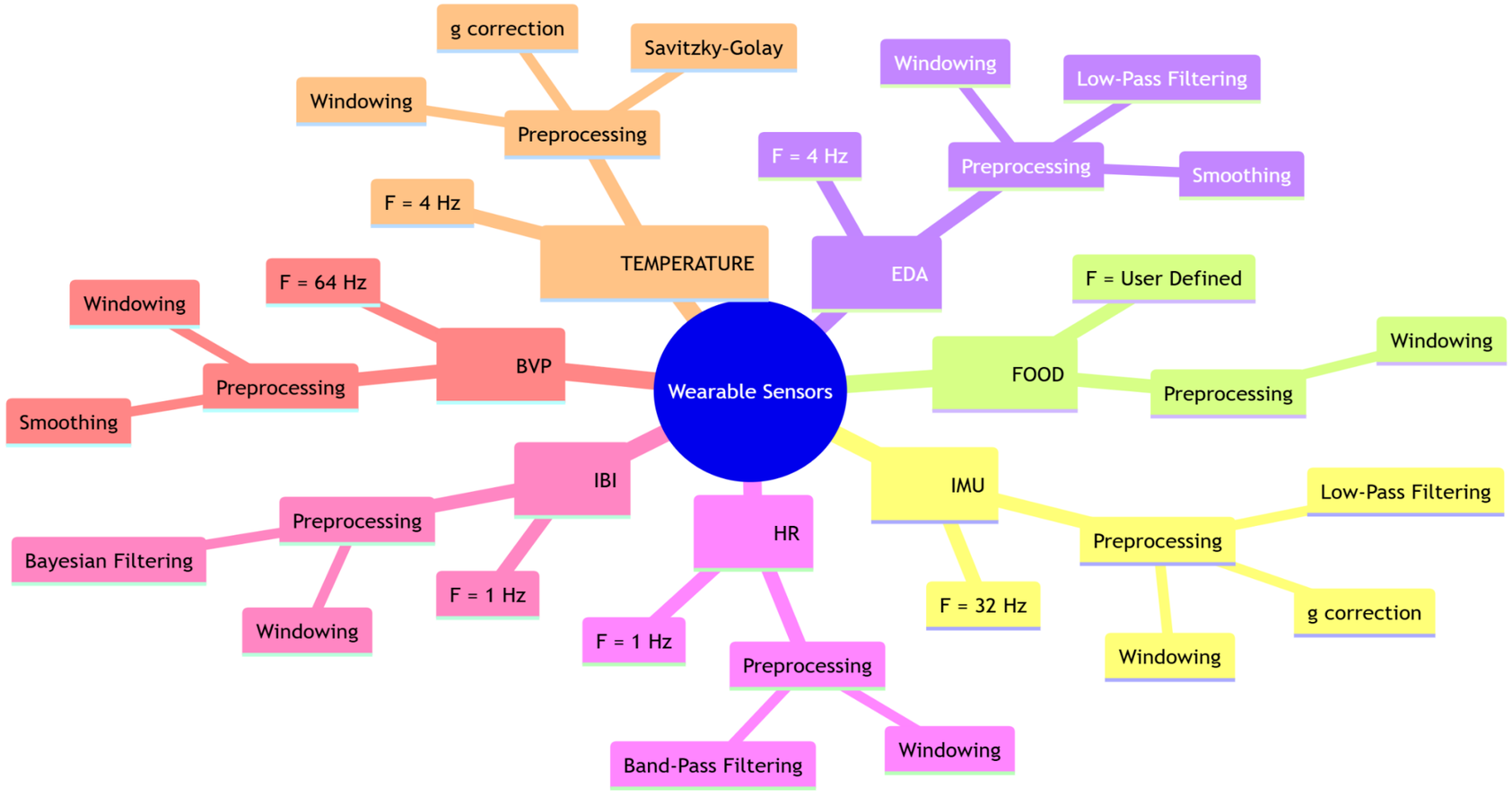


Figure 4.1: Preprocessing Steps performed on different sensors

Table 4.2: Mathematical Expressions for features used during this study

Feature	Description	Mathematical Expression
Biological Sex		
HbA1C	Glycated haemoglobin usually measured before the longitudinal data collection	
Mean of EDA, HR, IBI, (Temperature) T, and a (acceleration)	The mean of S (sensor value for the window for prediction usually equal to 5 min)	$\mu_S = \frac{\sum_{i=0}^{N} S}{N}$
Standard Deviation of EDA, HR, IBI, T, and a	The standard deviation of the S values for the length of the window	$\sigma_S = \sqrt{\frac{\left(\sum_{i=0}^{N} \mu_S - S\right)^2}{N}}$
Minimum Value of EDA, HR, IBI, T, and a		$\min_N S$
Maximum Value of EDA, HR, IBI, T, and a		$\max_N S$
First Quartile of EDA, HR, IBI, T and a	The value of 25% data point when the data are arranged in ascending order	$S(I)$ where *I* is the index of the S values in t ascending order rounded off to the nearest integer $\mathrm{I} = \frac{N+1}{4}$
Third Quartile of EDA, HR, IBI, T and a	The value of 75% data point when the data are arranged in ascending order	$S(I)$ where I is the index of the S values in ascending order rounded off to the nearest integer $\mathrm{I} = \frac{(N+1)x3}{4}$
Skewness of EDA, HR, IBI, T, and a	It is a measure of how symmetric the data are from the mean	$\mathrm{S} = \frac{\left(\sum_{i=0}^{N} \mu_S - S\right)^3}{(N-1)\sigma_S^3}$
Peak of EDA values	Peak of EDA values in the prediction window for peaks of prominence 0.3	$\sum_{i=0}^{N} P$
Rolling mean of two hours of EDA values		
Rolling sum of 2 h of EDA peaks		
Standard Deviation of IBI (SDNN)	It is a measure of HR variability	
Root mean square of successive differences of inter-beat interval (RMSSD)	It is a measure of heart rate variability that is related to autonomic nervous system tone	$\mathrm{RMSSD} = \sqrt{\frac{1}{\mathrm{N}-1}\sum_{\mathrm{n}=1}^{\mathrm{N}-1}[\mathrm{IBI}_{\mathrm{n}+1} - \mathrm{IBI}_{\mathrm{n}}]^2}$

Number of times IBI exceeds 50 ms (NN50)	It is a measure of HR variability and sympathetic nervous system activation	$NN_{50} = \sum_{i=1}^{n} a$ where a = 1 if $IBI_{i+1} - IBI_i > 50$ Else a = 0
pNN50	It is the measure of how many times in time N the IBI has exceeded 50 ms	$pNN_{50} = \frac{NN_{50}}{N}$
Calorie Sums	Rolling sums of 2 h, 8 h, and 24 h of calorie estimates are used	
Protein Sums	Rolling sums of 2 h, 8 h, and 24 h of protein consumption estimates are used	
Carbohydrate Sums	Rolling sums of 2 h, 8 h, and 24 h of carbohydrate consumption estimates are used	
Sugar Sums	Rolling sums of 2 h, 8 h, and 24 h of sugar consumption estimates are used	
Rolling mean of two-hour acceleration values	Used to estimate activity levels	
Rolling maximum value of two-hour acceleration values	Used to estimate activity levels	
Activity Bouts	When the mean of the window exceeds the rolling mean of acceleration values	
Individual Number	Used to model individuality	

The efficacy of the features was verified based on correlation and mutual information for discrete and t-Distributed Stochastic Neighbour Embedding (t-SNE) plots. Correlation is used to measure the independent features given in the supplementary materials Figure S5 . These features are used to train ML models.

DT is a non-parametric ML algorithm used for classification and regression tasks. It splits the data into subsets based on the value of the input features, forming a tree structure where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome. The goal is to create a model that predicts the target variable by learning simple decision rules inferred from the data features.

SVM finds the hyperplane that best separates the classes in the feature space, maximizing the margin between the closest points of the classes (support vectors). SVM is effective in high-dimensional spaces and is particularly useful for problems where the number of dimensions exceeds the number of samples.

RF is an ensemble learning method that combines multiple DTs to improve predictive performance and control overfitting. Each tree is trained on a random subset of the data and the features. The final prediction is made by averaging the outputs of individual trees (for regression) or by majority voting (for classification).

LDA is a dimensionality reduction technique used for classification. It projects the data onto a lower-dimensional space where the classes are most separable. LDA assumes that the features follow a Gaussian distribution and that different classes have identical covariances. It finds linear combinations of the features to predict the labels.

KNN is a simple, instance-based learning algorithm used for classification and regression. It predicts the target by finding the K training samples closest in distance to a new data point and returning the majority class (for classification) or the average value (for regression). KNN is non-parametric and makes predictions based on the entire dataset.

GNB is a probabilistic classifier based on Bayes' theorem, assuming independence between features given the class. It models the distribution of the features within each class as Gaussian.

LassoCV is a linear regression model that includes L1 regularization (lasso) to enforce sparsity, reducing the number of features by shrinking some coefficients to zero. The CV part stands for cross-validation, which is used to find the optimal regularization parameter. It helps prevent overfitting by penalizing the absolute size of the coefficients.

Ridge regression is a linear regression model with L2 regularization, which penalizes the squared magnitude of the coefficients. This regularization helps to prevent overfitting by shrinking the coefficients towards zero but unlike lasso, it does not enforce sparsity. It is useful when dealing with multicollinearity or when the number of predictors exceeds the number of observations.

AdaBoost, short for Adaptive Boosting, is an ensemble learning technique that combines multiple weak classifiers to create a strong classifier. It works by iteratively training classifiers on weighted versions of the data, where the weights are adjusted to focus on the hardest-to-classify samples. The final model is a weighted sum of the individual classifiers.

XGBoost (extreme gradient boosting) is an efficient and scalable implementation of the gradient boosting framework. It builds an ensemble of DT in a sequential manner, where each tree corrects the errors of its predecessor. XGBoost uses advanced regularization techniques to reduce overfitting and includes features like parallel tree construction, handling missing values, and optimized computations.

For both regression and classification, numerical features were normalized using the Z-score. The Z-score is defined in equation 4.1.

$$z = \frac{x - \mu}{\sigma} \qquad 4.1$$

where x gives the value of the feature, μ is the mean of the feature's distribution, and σ is its standard deviation. There are two categorical features in this dataset: participant ID and gender. One-hot encoding is a method used to transform categorical data into a numerical format suitable for ML models. It converts each category into a unique binary vector, where only one element is set to 1 and the rest are 0. In this work, the categorical features are one-hot encoded. The ML models are trained using a subset of data called training data. Features from an independent subset called validation data are used to predict IG values; these values are compared with actual IG values for determining the performance of the model.

The following performance metrics are used in the comparison of the regression models.

- MAE: MAE measures the average absolute difference between the predicted and actual IG values. It is less sensitive to outliers.
- RMSE: It is the root of the average squared difference between the predicted and actual IG values.
- MAPE: MAPE measures the average ratio of error (difference between the actual and predicted value) with the actual IG value.

- R-squared (R^2) and Adjusted R-squared: R^2 measures the proportion of variance in the CGM values explained by the input smart watch features. A higher R^2 indicates a better fit of the model to the data.
- MSLE: It is the average difference between the log of the actual and predicted IG values. It is specifically less sensitive to outliers.

For classification, the following parameters are used to define the performance of the models. A true positive (TP) occurs when the model correctly predicts a positive class for an instance that is actually positive. A true negative (TN) is when the model correctly predicts a negative class for an instance that is actually negative. A false positive (FP), also known as a Type I error, happens when the model incorrectly predicts a positive class for an instance that is actually negative. Conversely, a false negative (FN), or Type II error, occurs when the model incorrectly predicts a negative class for an instance that is actually positive.

The following performance metrics are used to compare the classification models.

- Accuracy (%): Accuracy measures the proportion of correctly classified instances out of the total instances. It is given by (TP+TN/TP+TN+FP+FN);
- Precision: Precision, also known as positive predictive value, measures the proportion of true positive predictions among all positive predictions made by the model. It is given as (TP/TP+FP).
- Recall: Recall, also known as sensitivity or the true positive rate, measures the proportion of true positives identified by the model out of all actual positives. It is given as (TP/TP+FN).
- F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall.
- ROC (Receiver Operating Characteristic) Curve: The ROC curve plots the true positive rate (recall) against the false positive rate at various threshold settings. The ROC AUC (Area Under the Curve) measures the model's ability to discriminate between classes, with a higher AUC indicating a better performance.

To make sure that the difference between performance metrics is statistically significant, they are compared using the Friedman test. This is done by dividing the dataset into 10 folds, training it on 9 folds and testing on the remaining fold (once for each fold). The performance metrics are recorded, and these differences are interpreted using the Nemenyi post hoc test.

4.5 Results

4.5.1 Feature Calculation

These features are calculated in python using NumPy (Harris et al., 2020), pandas (McKinney & Team, 2015), and JAX. Notably, some features, such as those related to proteins and carbohydrates over different time windows, exhibit strong correlations within their groups. Conversely, features like heart rate variability (HRV) metrics and activity measures demonstrate more independence, as indicated by their lighter shades in Figure 4.2. Some of the features appear uncorrelated with the target variable as the correlations capture linear relationships and not their complex interactions.

The stronger shades of red signify a positive correlation, and blue signifies a negative correlation. The lighter shades signify the features that have a smaller correlation, meaning that they are potentially independent. Figure 4.3 illustrates the Pearson correlation coefficients between various features and IG values. Each feature's correlation with the dependent variable is visually represented, where positive correlations extend to the right and negative correlations to the left. Features such as activity counts, historical accelerometer data, HR metrics, and carbohydrate intake exhibit high degrees of correlation with the CGM values.

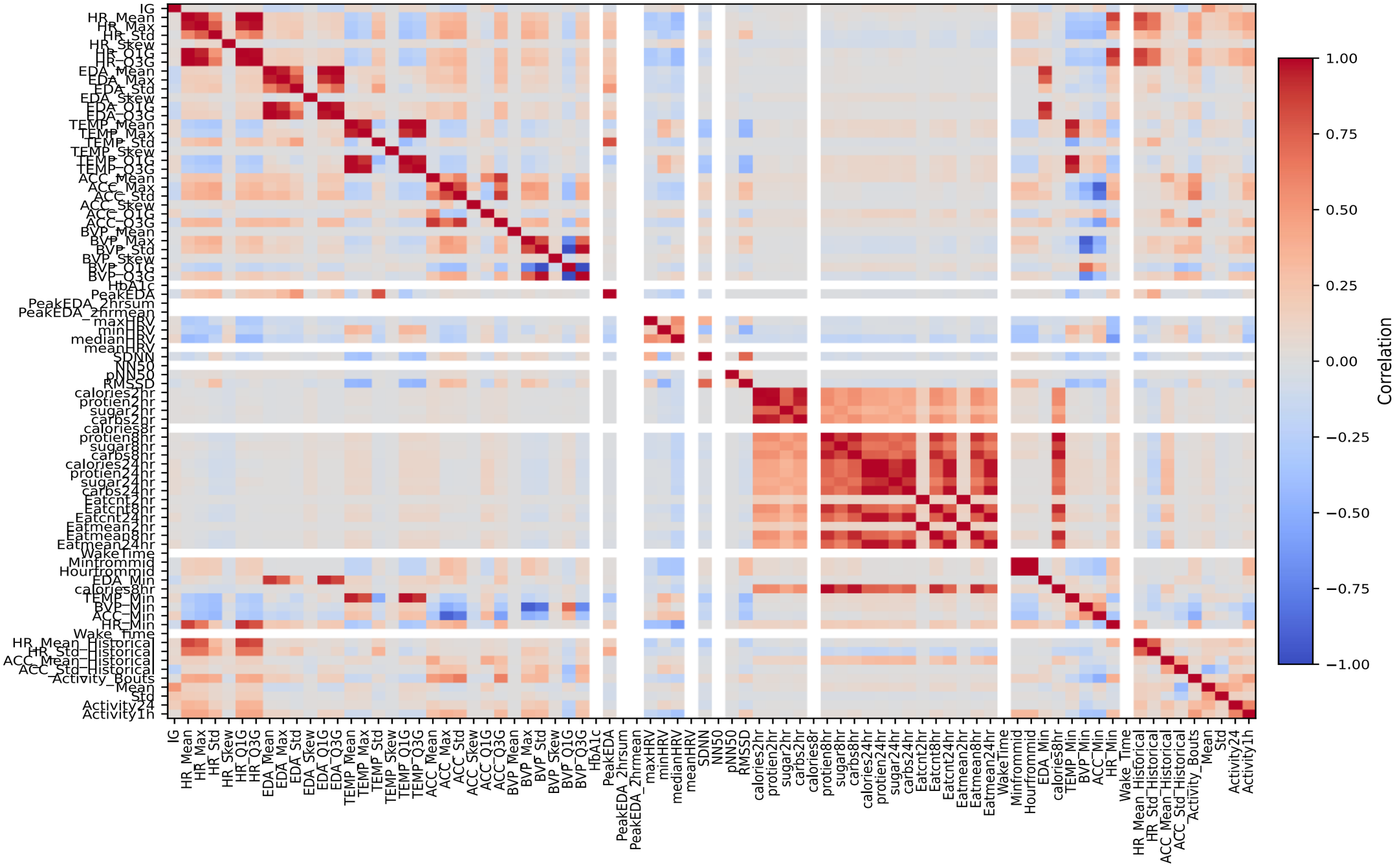


Figure 4.2: Correlation between different features

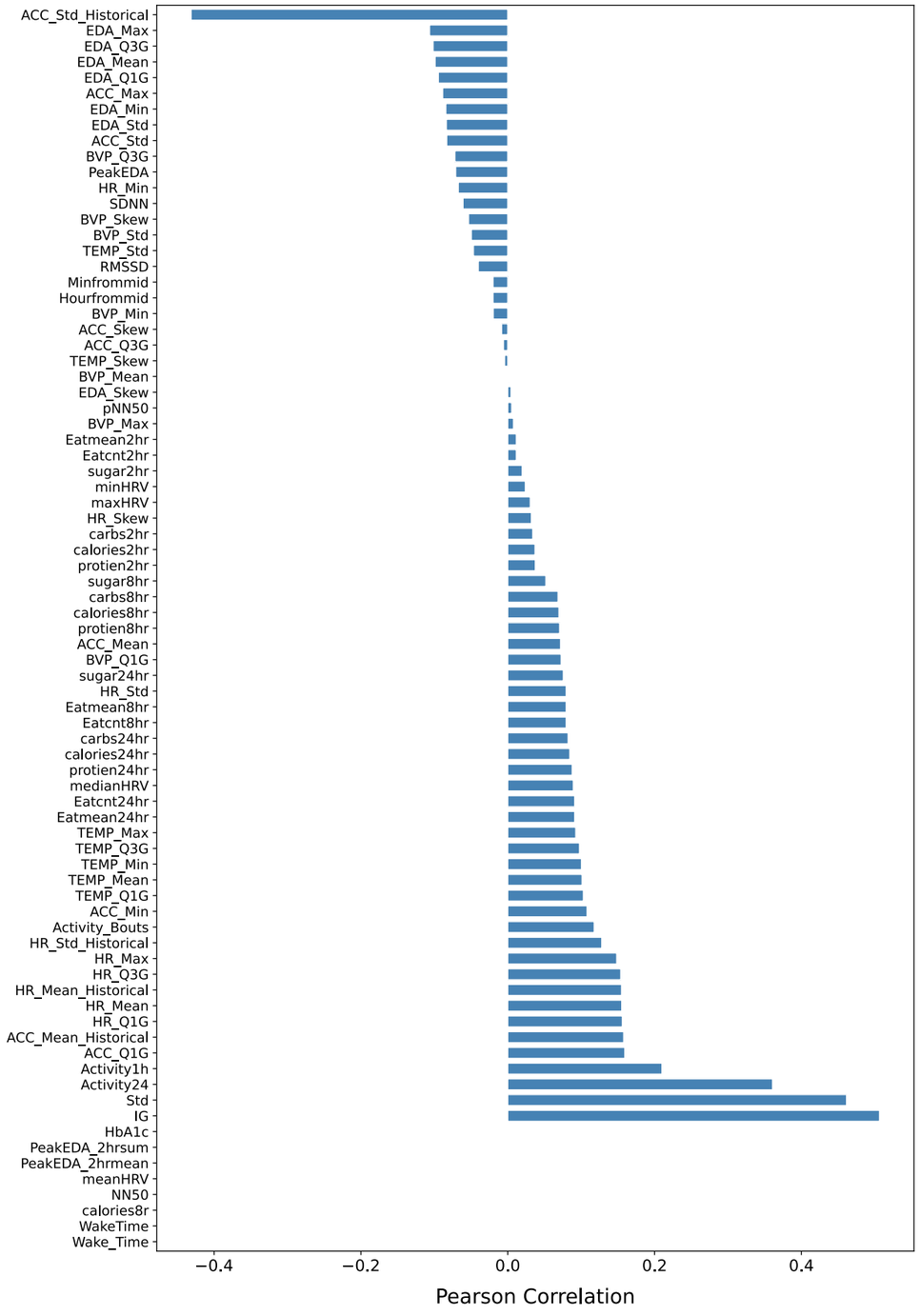


Figure 4.3: Feature correlations with IG

4.5.2 Regression

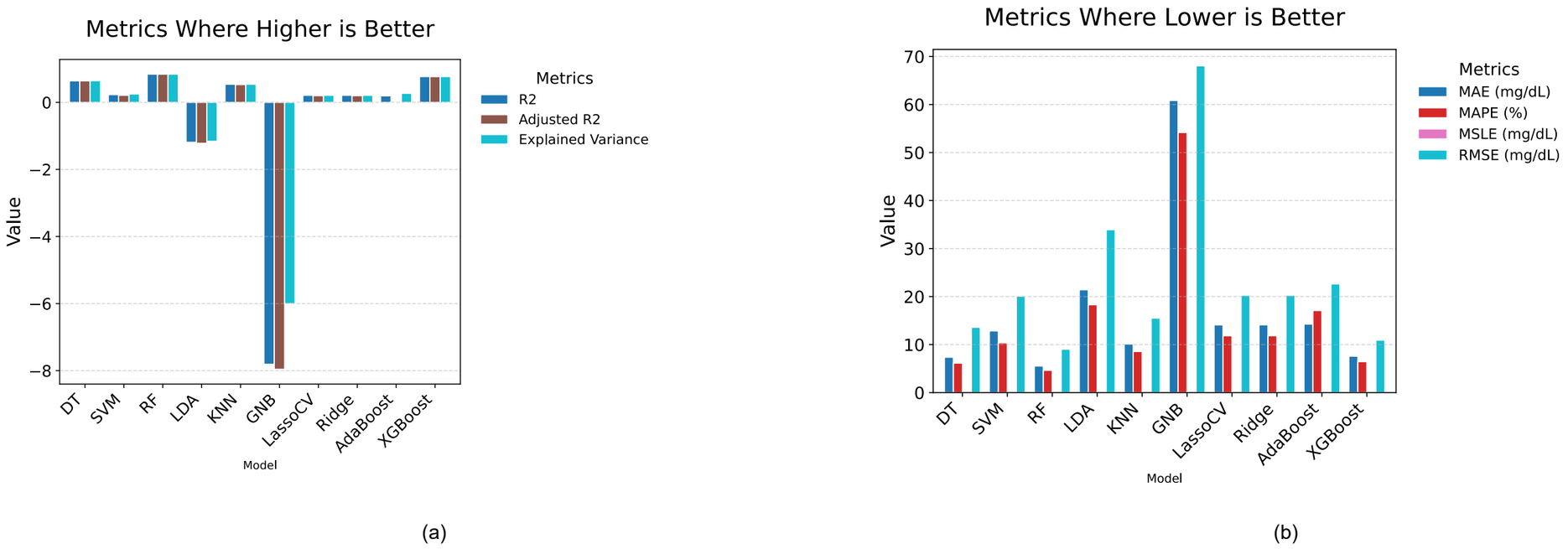
Regression models predict IG values based on smart watch and food log features. These ML models are used for IG prediction: DT, SVM, RF, LDA, KNN, GNB, LassoCV, Ridge, AdaBoost, and XGBoost. After identifying the best performing model type (RL), its hyperparameters were tuned using Bayesian Optimization.

The feature data were split into (70%) training and (30%) testing. The models were trained on the training set. The trained models were used to predict IG values on the unseen testing data. The predicted IG values were subtracted from the actual IG value to determine the *error* value. The error values were used to calculate various performance metrics of the regression models. Table 4.3 compares the performance metrics of the ML models in predicting IG values based on the input features.

Table 4.3: Performance of different regression models

Model	MAE (mg/dL)	MAPE (%)	R^2	Adjusted R^2	MSLE (mg/dL)	Explained Variance	RMSE (mg/dL)
DT	7.379	6.15	0.64	0.64	0.0115	0.6475	13.61
SVM	12.86	10.37	0.23	0.21	0.004	0.25	20.09
RF	5.54	4.65	0.84	0.84	0.068	0.84	9.04
LDA	21.42	18.30	−1.19	−1.22	0.014	−1.16	33.94
KNN	10.14	8.57	0.54	0.53	0.06	0.54	15.51
GNB	60.85	54.18	−7.81	−7.96	0.25	−6.00	68.07
LassoCV	14.11	11.84	0.21	0.20	0.02	0.21	20.27
Ridge	14.12	11.85	0.21	0.20	0.025	0.21	20.27
AdaBoost	14.28	17.10	0.194	0.007	0.034	0.27	22.64
XGBoost	7.59	6.45	0.77	0.768	0.007	0.77	10.93

Figure 4.4 presents the performance parameters of the models visually. RF consistently has high performance across all metrics. The RF model has a high R^2, adjusted R^2, and explained variance. while maintaining a low MAE, MAPE, MSLE, and RMSE.



(a) (b)

Figure 4.4: Comparison of the performance metrics of regression models: (a) Bar chart comparing regression models on metrics where higher values denote better performance. Shown are the coefficient of determination (R^2), the adjusted R^2, and the explained variance for each model—higher bars indicate models that explain more variance in the glucose predictions.; (b) Bar chart comparing regression models on metrics where lower values denote better performance. Shown are mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared log error (MSLE), and root mean squared error (RMSE) for each model—shorter bars indicate more accurate glucose predictions.

In both panels of Figure 4.4, the Gaussian Naive Bayes model trails the others. This is likely because GNB relies on the strong assumption that all predictors are independent, yet blood glucose levels are driven by complex, interdependent physiological and behavioural factors.

To make sure that the difference between the metrics for each model is significant, a Friedman test is carried out for each metric. An example of this analysis is shown here for reference. The Friedman statistic for values of MAE for all the models for all folds is 70.0, with ($p = 1.47 \times 10^{-12}$) showing that the difference is significant.

The Nemenyi post hoc test results (Figure 4.5) for mean absolute error (MAE) reveal significant and non-significant differences in model performance. Significant differences ($p < 0.05$) were observed between AdaBoost and DT ($p = 0.001$), AdaBoost and KNN ($p = 0.006$), AdaBoost and Random Forest ($p = 0.001$), AdaBoost and XGBoost ($p = 0.001$), DT and LassoCV ($p = 0.001$), DT and Ridge ($p = 0.006$), Random Forest and LassoCV ($p = 0.001$), Random Forest and Ridge ($p = 0.001$), SVM and Random Forest ($p = 0.006$), and XGBoost and LassoCV ($p = 0.006$). No significant differences ($p \geq 0.05$) were found between AdaBoost and LassoCV, AdaBoost and Ridge, AdaBoost and SVM, DT and KNN, DT and Random Forest, DT and SVM, DT and XGBoost, KNN and LassoCV, KNN and Random Forest, KNN and Ridge, KNN and SVM, KNN and XGBoost, LassoCV and Ridge, LassoCV and SVM, Random Forest and XGBoost, Ridge and SVM, Ridge and XGBoost, and SVM and XGBoost. Nemenyi Post Hoc Analysis for all the performance metrics are given in the supplementary material.

The parameters of the RF model are tuned using Optuna (Akiba et al., 2019). Figure 4.6 shows the optimization process. The number of estimators is the number of DT in the RF model, maximum depth is the maximum depth of each DT, and minimum sample leaf is the smallest number of samples that should be present in the leaf node after splitting a node. This hyperparameters are reported in Table 4.4. The objective value for this tuning process is negative mean absolute error (MAE).

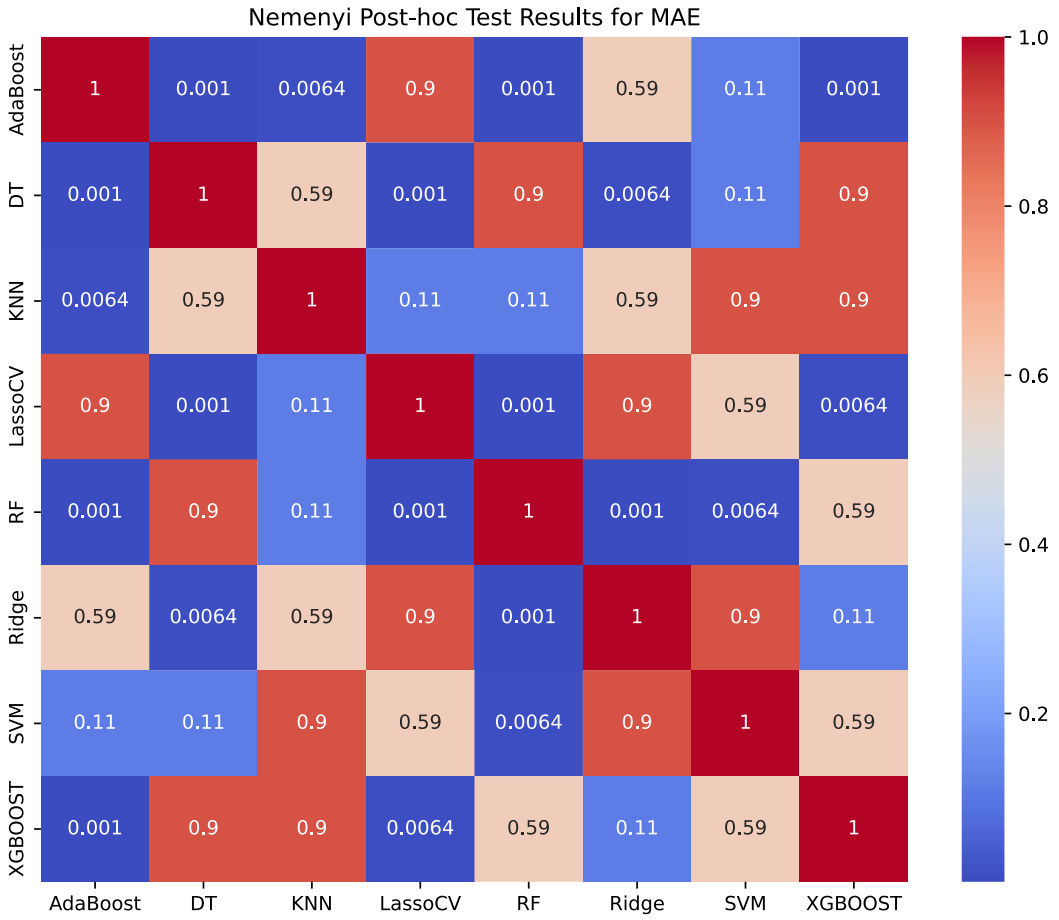
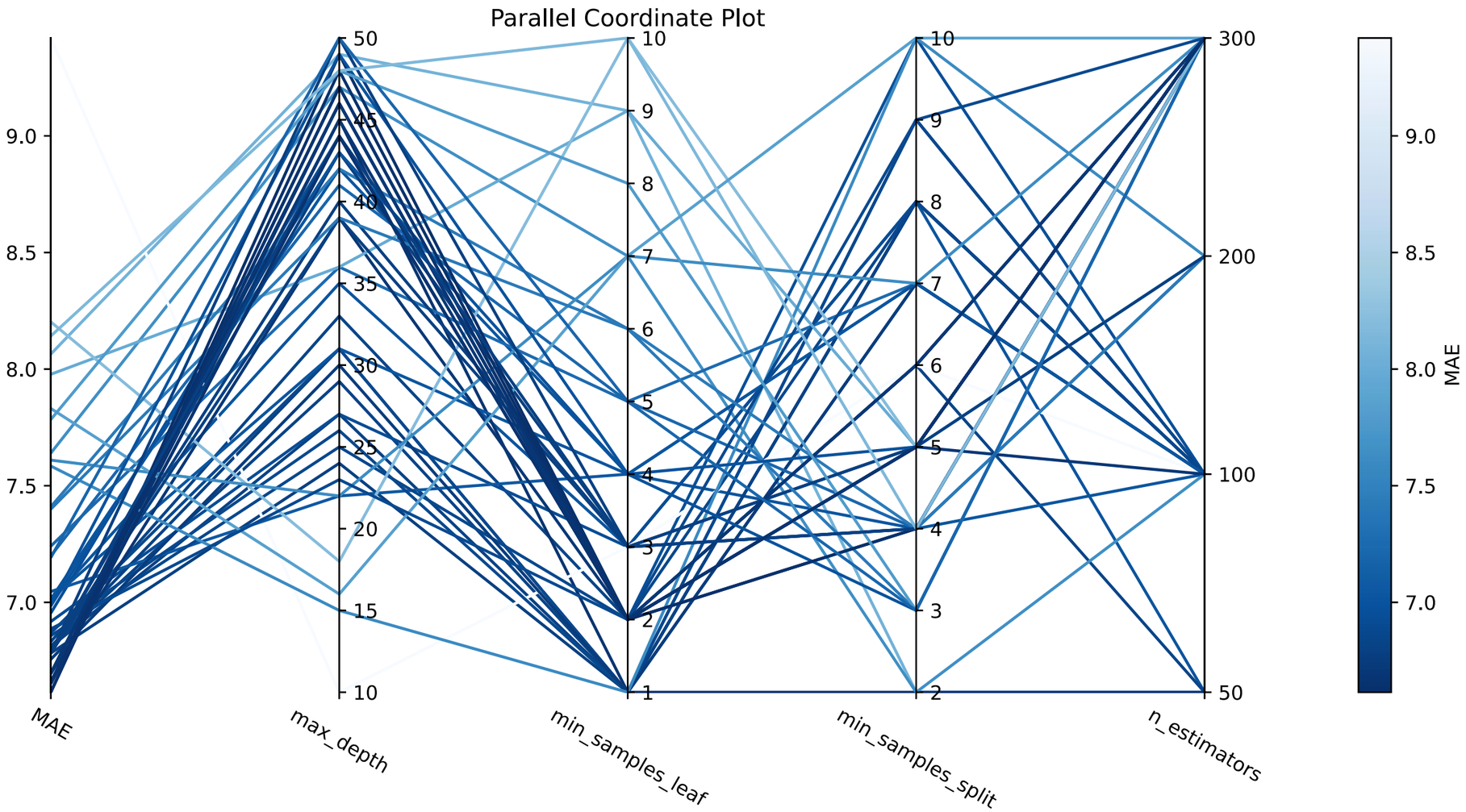


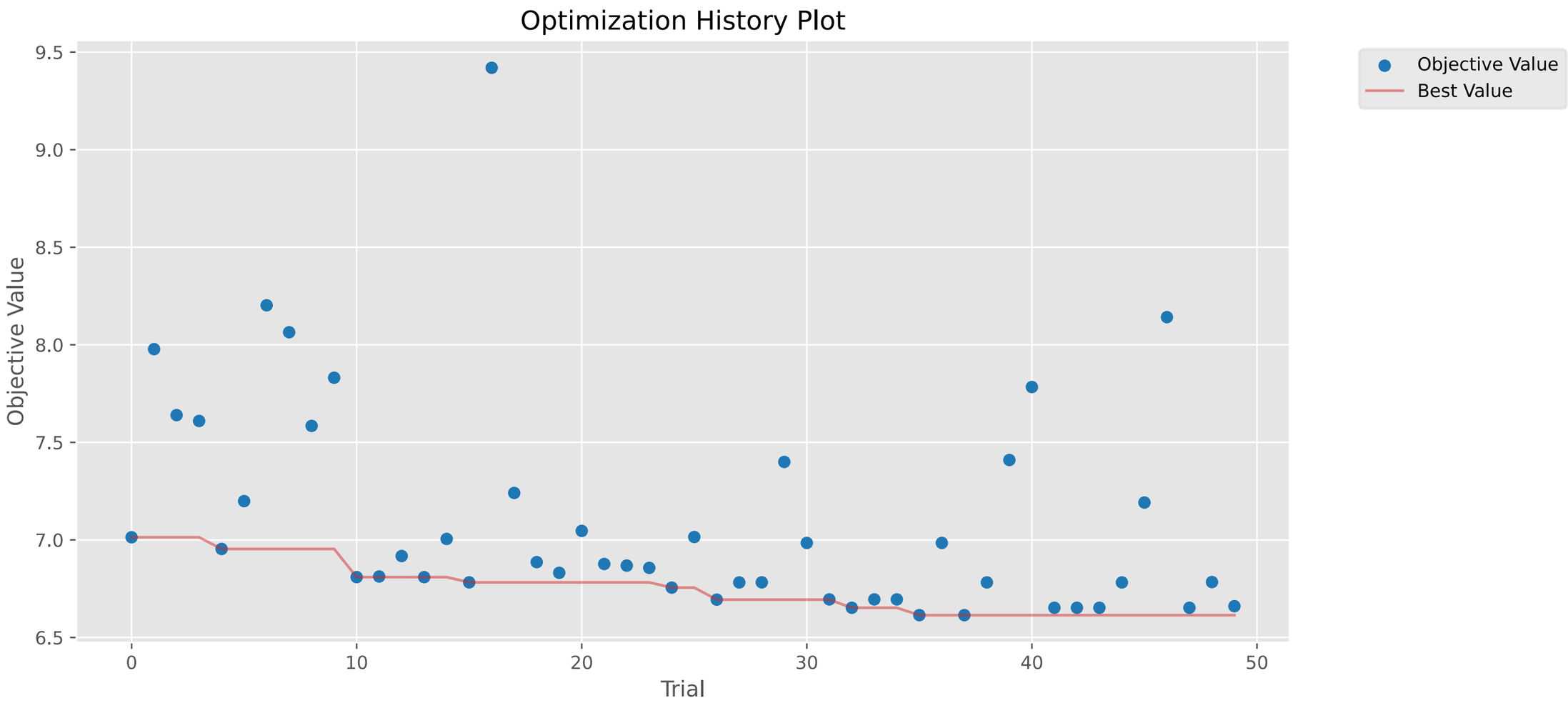
Figure 4.5: Nemenyi post hoc analysis of the Friedman test for MAE across all the models.

Table 4.4: Hyperparameters of the best performing model

Hyperparameter	Number of Estimators	Maximum Depth	Minimum Sample Split	Minimum Leaves Per Sample
Value	178	26	10	1


Parallel Coordinate Plot
MAE
max_depth
min_samples_leaf
min_samples_split
n_estimators
MAE

(a)



(b)

Figure 4.6: Bayesian Optimization for hyperparameter tuning: (a) Parallel coordinates shaded with the objective value; the objective for the optimization is the RMSE value. (b) The evolution of the RMSE over the number of iterations.

4.5.3 Classification

The classification models categorize IG values into normal, high, and low. Most IG values belong to the normal class in this dataset. To overcome this class imbalance, the number of samples per class is stratified by down sampling the normal class to 2500 samples. The total number of samples is 7500 (2500 samples per class). A total of 70% of the samples are used in training the ML models, while 30% are used for testing the performance of those models (RF, DT, SVM, LDA, KNN, GNB, Ridge, AdaBoost and XGBoost). After identifying the best performing model type (RL), its hyperparameters are tuned using Bayesian Optimization.

The classification models are compared based on the relevant performance metrics (recall, precision, accuracy, f1-score). These results are displayed visually in Figure 4.7 (a) and (b)

To make sure that the difference between the metrics for each model is significant, a Friedman test is carried out for each metric. An example of this analysis is shown here for reference. The Friedman statistic for values of accuracy for all the models for all folds is 78.33 with (p = 1.05 × 10^{-13}), showing that the difference is significant. To understand for which models the comparison is significant, a Nemenyi post hoc (Figure 4.8) analysis for the results is conducted and plotted as a heatmap. Nemenyi Post hoc analysis for all the performance metrics is given in the supplementary material.

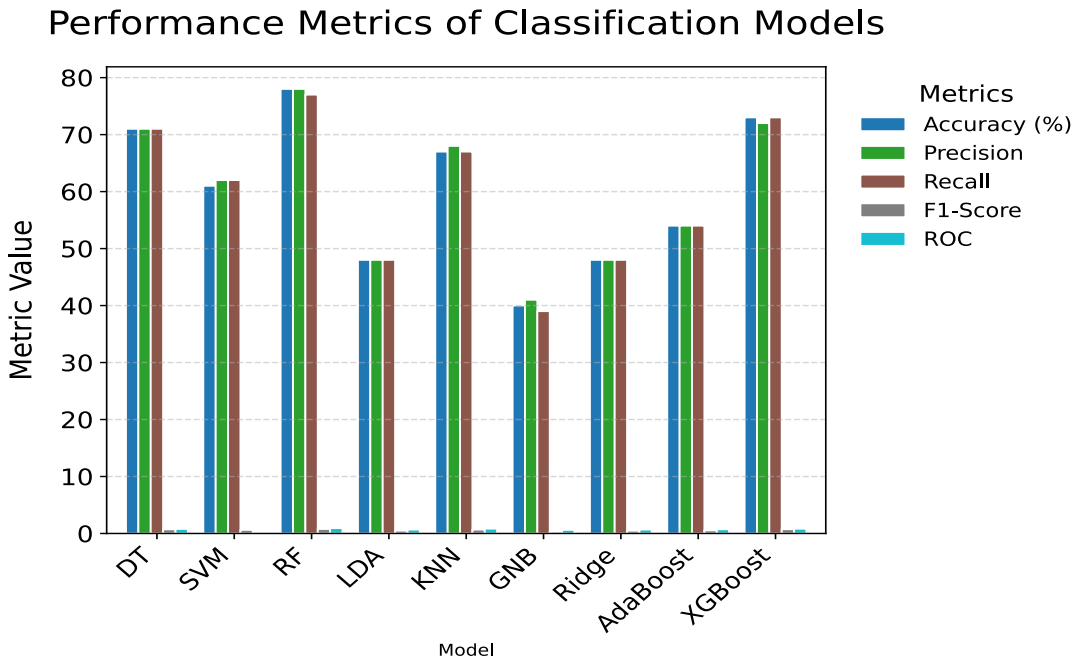


Figure 4.7: Comparison of the performance metrics of classification models: (a) Normalized spider plot for different performance metrics of classification; (b) bar plot for performance measures of different models.

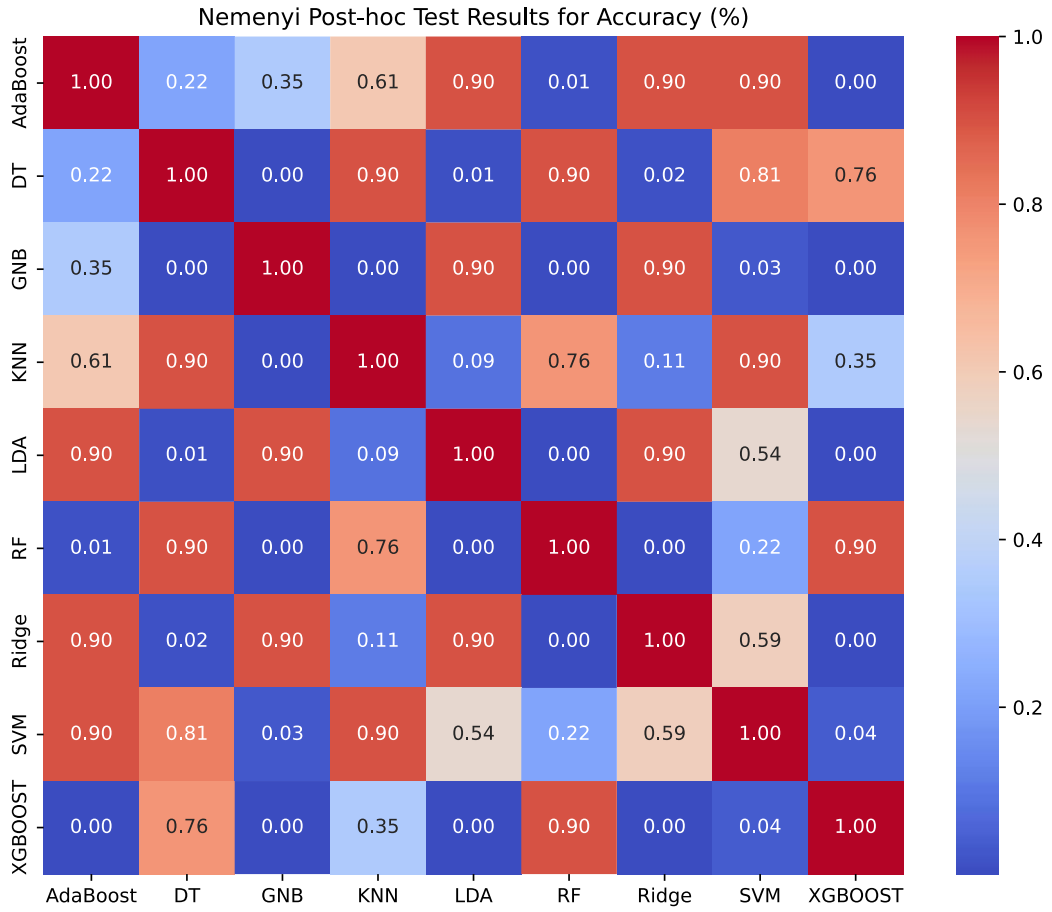


Figure 4.8: Nemenyi post hoc test results for accuracy (%).

The Nemenyi post hoc test results for accuracy reveal significant and non-significant differences in model performance. Significant differences ($p < 0.05$) were observed between AdaBoost and XGBoost ($p = 0.001$), DT and GNB ($p = 0.001$), GNB and KNN ($p = 0.006$), Random Forest and GNB ($p = 0.001$), GNB and XGBoost ($p = 0.001$), LDA and RF ($p = 0.002$), and LDA and XGBoost ($p = 0.001$). No significant differences ($p \geq 0.05$) were found between AdaBoost and DT, AdaBoost and GNB, AdaBoost and KNN, AdaBoost and LDA, DT and KNN, DT and Random Forest, DT and SVM, DT and XGBoost, KNN and LassoCV, KNN and Random Forest, KNN and Ridge, KNN and SVM, KNN and XGBoost, LassoCV and Ridge, LassoCV and SVM, Random Forest and XGBoost, Ridge and SVM, and SVM and XGBoost.

The performance of the models across different performance metrics is shown in Table 4.5. RF outperformed the other models across all the performance metrics.

Based on these results, RF outperforms other models in the classification tasks as well. In light of this the hyperparameters of the RF model are optimized using Bayesian Optimization process to find the most optimal hyperparameters.

Table 4.5: Performance metrics of classification models.

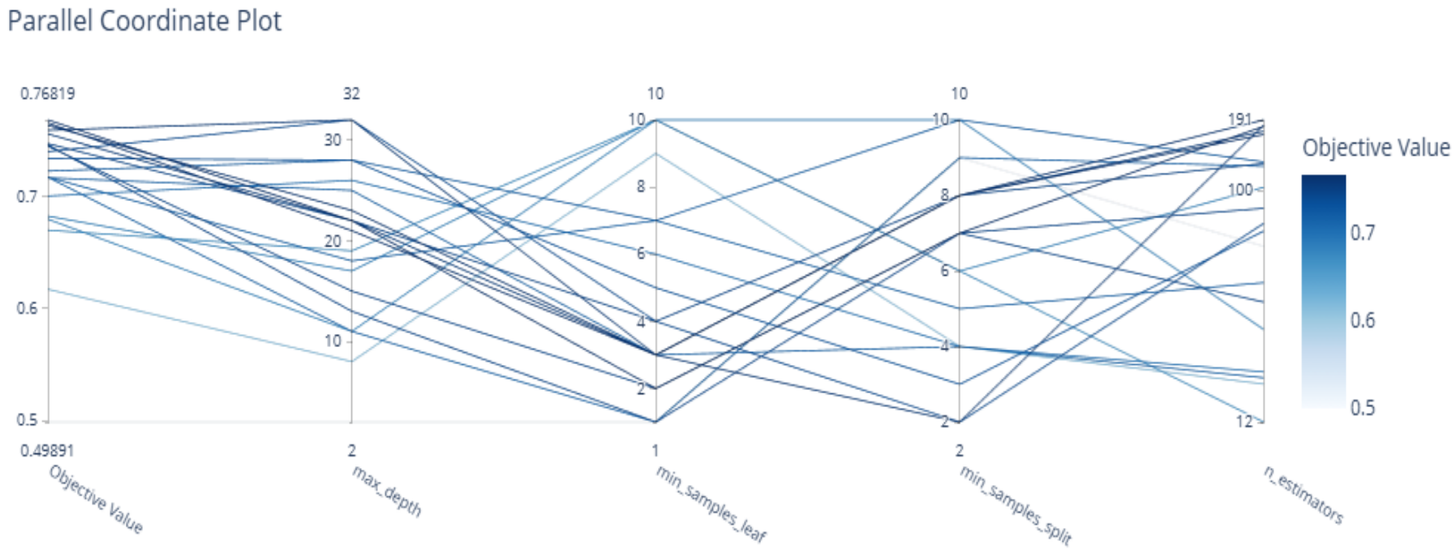
Model	Accuracy (%)	Precision	Recall	F1-Score	ROC
DT	71	71	71	0.71	0.78
SVM	61	62	62	0.61	0.25
RF	78	78	77	0.77	0.92
LDA	48	48	48	0.48	0.67
KNN	67	68	67	0.66	0.83
GNB	40	41	39	0.31	0.61
Ridge	48	48	48	0.47	0.67
AdaBoost	54	54	54	0.53	0.72
XGBoost	73	72	73	0.73	0.82

The optimization process is displayed in Figure 4.9 and the results of this optimization are displayed in Table 4.6. The objective value for this optimization is prediction accuracy of the RF model.

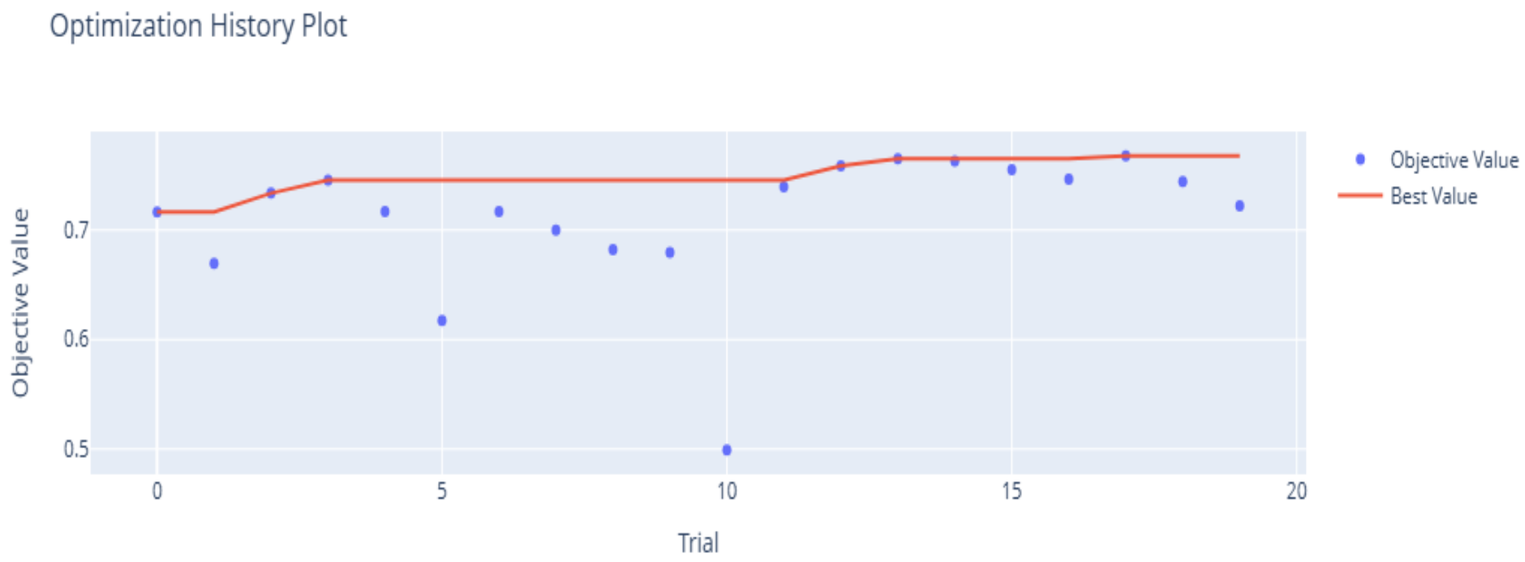
Table 4.6: Optimal hyperparameters measured using Bayesian Optimization using Optuna

Hyperparameter	Number of Estimators	Maximum Depth	Minimum Sample Split	Minimum Leaves Per Sample
Value	130	22	7	2

The performance of tuned RF model for glucose level classification is further explored to verify the effects of tuning the model. The results include plots such as class prediction error, which displays the correct and misclassifications for each class as a bar length, confusion matrix, which displays the extent to which one class may be misclassified as another class. The receiver operator curve, which is a plot between true positive rates and false positive rates. Thes plots are shown in Figure 4-10. These plots show the superior performance of RF model in discerning the personalized high, low and normal glucose levels.

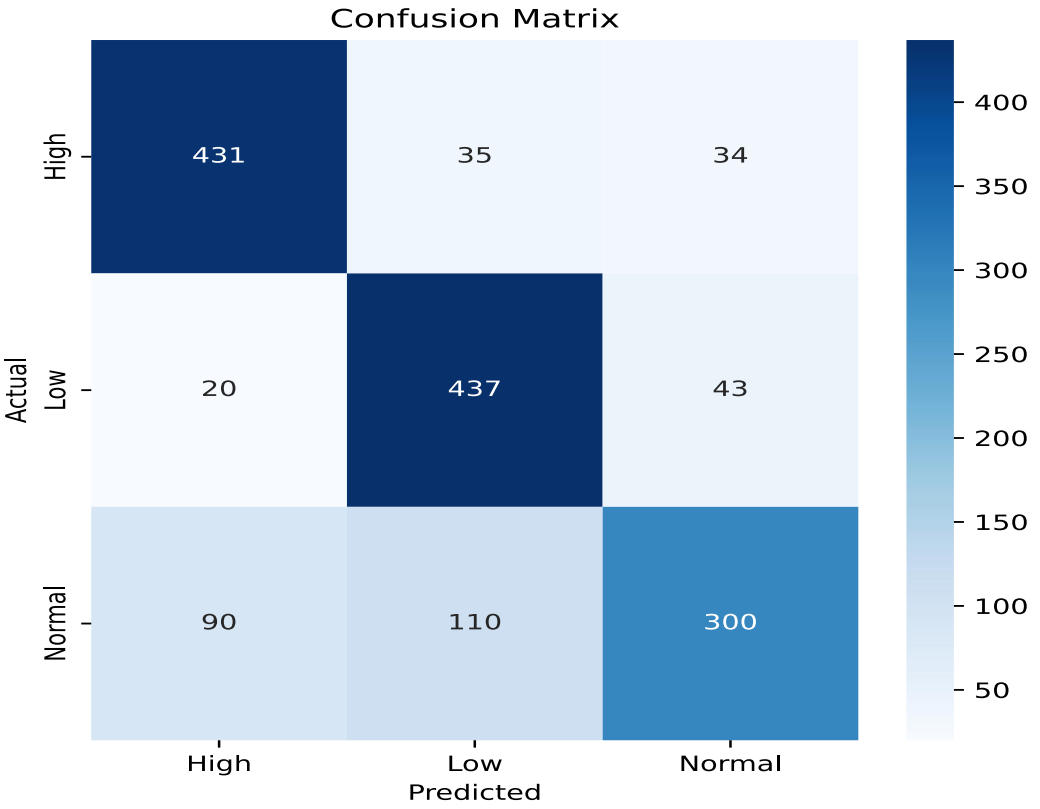


(**a**)

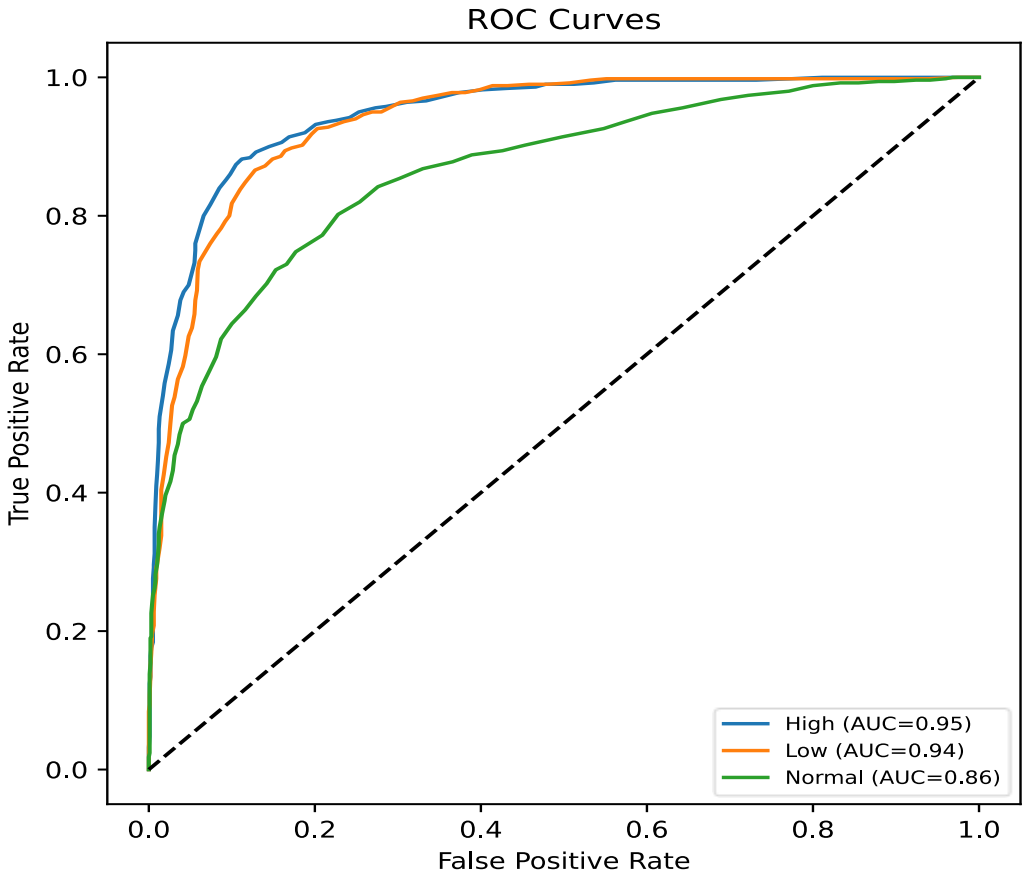


(**b**)

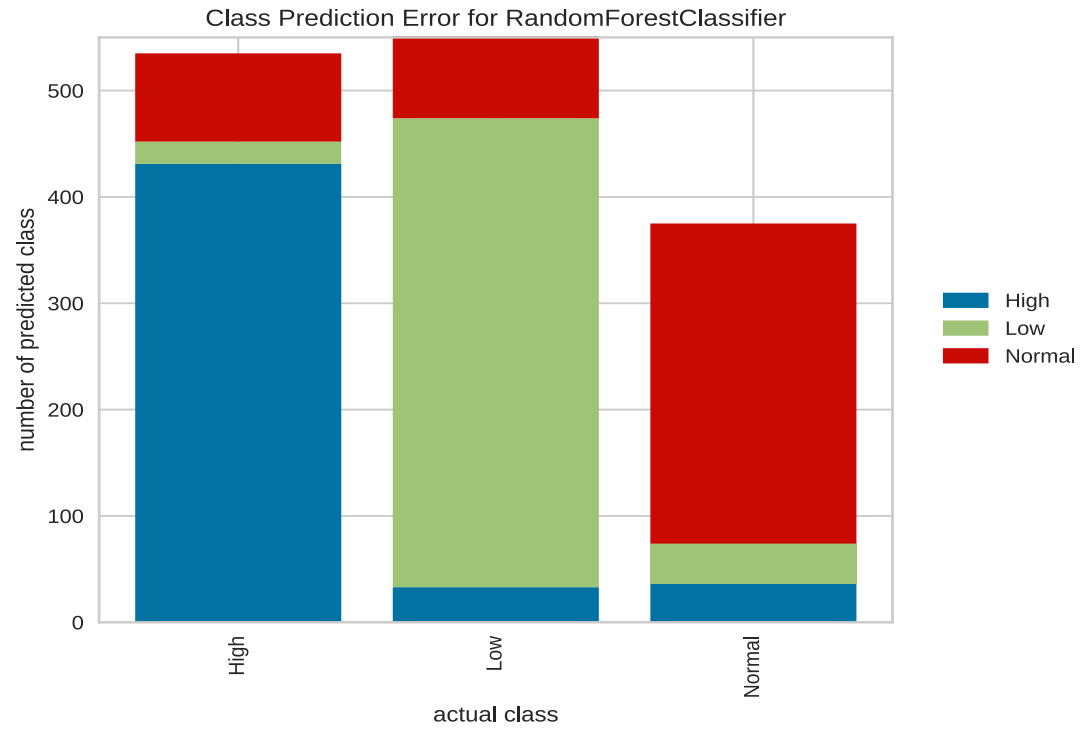
Figure 4.9: Bayesian Optimization for hyperparameter tuning: (a) Parallel coordinates shaded with the objective value; the objective for the optimization is accuracy. (b) The evolution of the accuracy over the number of iterations


Confusion Matrix
High
Low
Normal
Actual
431
35
34
20
437
43
90
110
300
400
350
300
250
200
150
100
50
High
Low
Normal
Predicted

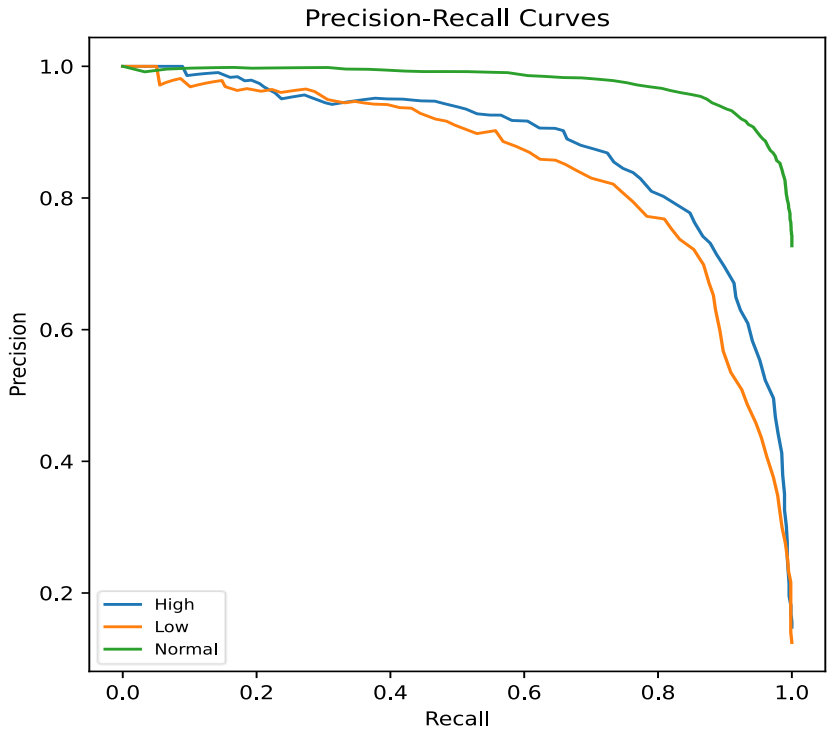
(a)


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
High (AUC=0.95)
Low (AUC=0.94)
Normal (AUC=0.86)

(b)



(c)



(d)

Figure 4.10: Performance of the tuned Random Forest model on validation data of the balanced dataset: (a) Confusion matrix of the tuned RF classifier for validation data of the balanced dataset, (b) ROC curves of the tuned RF classifier for validation data of the balanced dataset, (c) class prediction error of the tuned RF classifier for validation data of the balanced dataset, and (d) precision recall curve of the tuned RF classifier for validation data of the balanced dataset. Nemenyi post hoc analysis of performance metrics of classification models are given in Supplementary Material.

4.5.4 Model Explanations

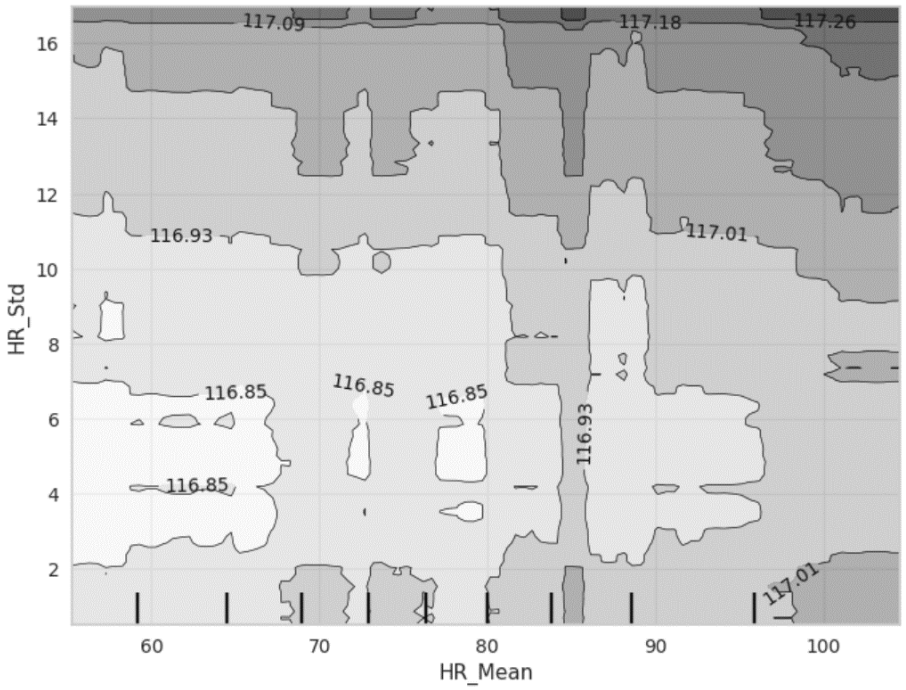
The best regression and classification models are explained. To better understand why tree models, perform better than kernel-based models such as SVMs, generative models such as GNB, and non-parametric models such as KNNs, partial dependence plots are used to show the complex interaction of the features modelled.

According to the literature, tree-based models are robust to noise and suitable for visualizing feature interactions. Here, we plot the partial dependence plots of two features we believe interact with each other in a complex manner (HR_Mean and HR_Std). As can be seen from the partial dependence plot (PDP) from the RF and the LDA in Figure 4.11, the PDP of the RF model represents a complex relationship, whereas for the linear model, the PDP shows a linear relationship, resulting in lower performance metrics for regression.

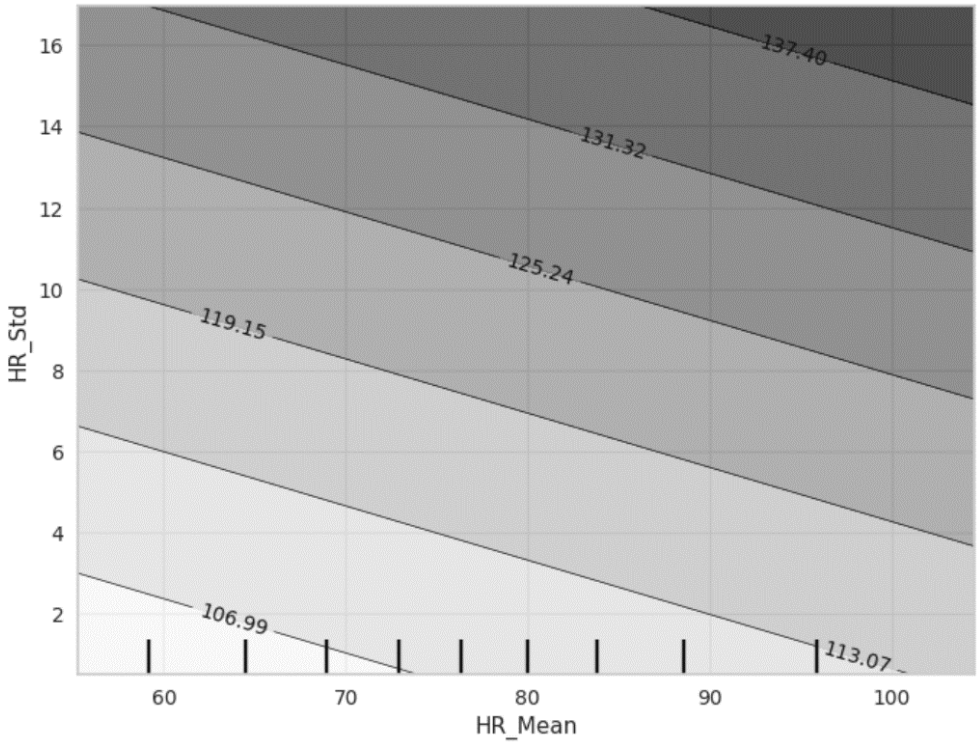
The SHAP plot in Figure 4.12a reveals the distribution and impact of features on the RF model's predictions for classifying CGM values into high, low, and normal categories. Each feature's influence is illustrated by the spread of dots along the *x*-axis, indicating the SHAP values. A wider spread of dots signifies a greater variance in the feature's impact across different data points. For instance, Hourfrommid and HR_Mean have a broad range of SHAP values, showing they can significantly sway the predictions towards both high and low CGM classes. The colour gradient of the dots, from blue (low feature value) to red (high feature value), further elucidates how different levels of a feature affect the prediction. This spread and colour coding collectively depict the nuanced contributions of each feature, demonstrating the complex interplay between physiological metrics, food intake, and glucose levels. This shows that for the classification problem, the circadian features are relatively more important than other features. This is consistent with earlier works (Ali et al., 2023, 2024; Bent, Cho, Henriquez, et al., 2021).

Figure 4.12b demonstrates how each feature contributes to the regression model's prediction of CGM values, highlighting the importance and influence of physiological metrics and activity levels captured by the Empatica E4 smart watch, as well as historical data. Key features like '*PeakEDA*', '*HR_Mean_Historical*', and '*ACC_Std_Historical*' exhibit a broad range of SHAP values, indicating significant variability in their impact on CGM predictions. For instance, high values of '*PeakEDA*' tend to push predictions higher, while low values often push predictions lower. Similarly, '*HR_Mean_Historical*' consistently shows a positive impact on CGM values, with high feature values leading to higher predictions. The varied influence of '*ACC_Std_Historical*' and 'Activity24' further underscores the complexity and interplay of factors affecting glucose levels. This

comprehensive visualization of feature contributions provides valuable insights into the model's decision-making process.

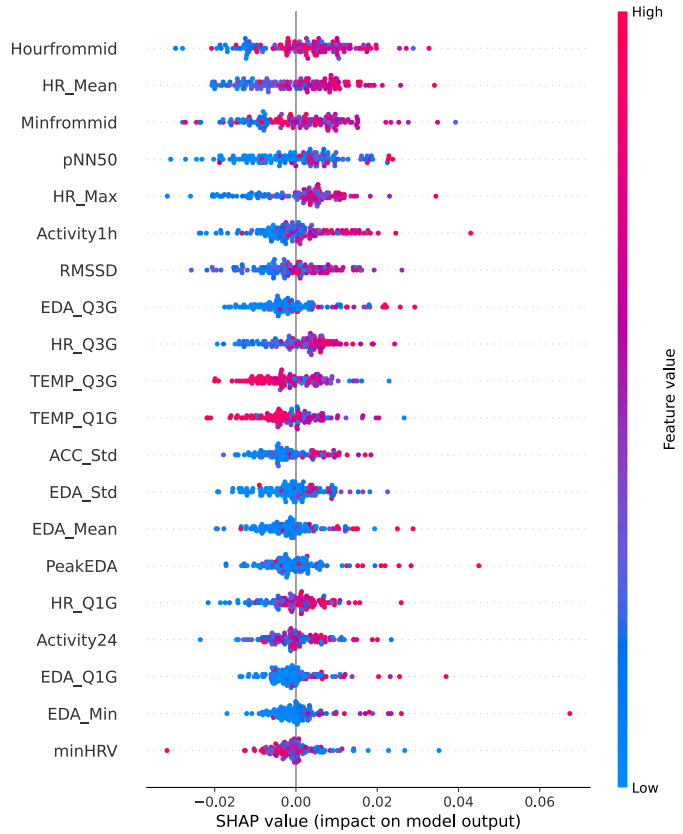


(a)

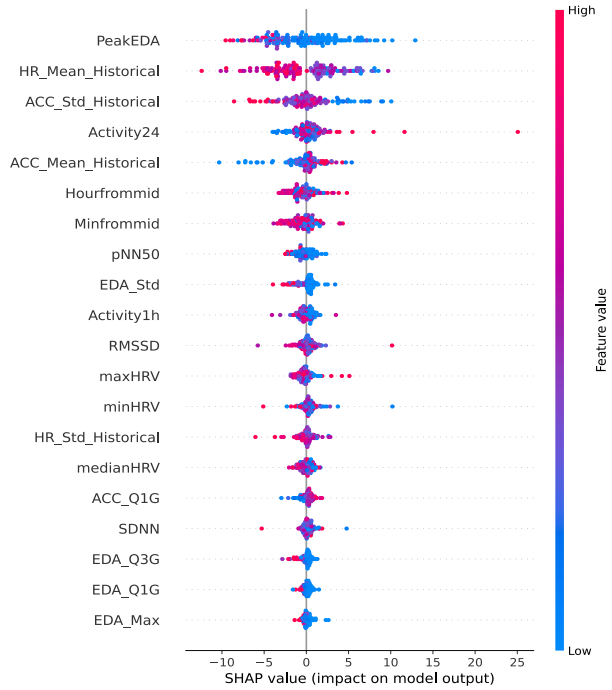


(b)

Figure 4.11 : Comparison of PDP plots for standard deviations of heart rate and mean heart rate: (a) The RF PDP captures a complex relationship, resulting in a higher accuracy; (b) the LDA assumes a linear relationship, resulting in a lower performance. The stronger shades in the PDP represent higher IG predicted



(a)



(b)

Figure 4.12: SHAP summary plots for classification and regression. (a) SHAP values for classification, (b) SHAP values for regression.

4.6 Discussion

For both classification and regression tasks RF, has a superior performance, while the other tree-based model, DT, is not that far behind. XGBoost, which is also a tree model,

performs well in both the tasks as well. Tree models are known to perform well in cases when there are nonlinear relationships. KNNs (Cover & Hart, 1967) and tree-based models (RF, DT, and XGBoost) are both equipped to handle nonlinear relationships between the data.

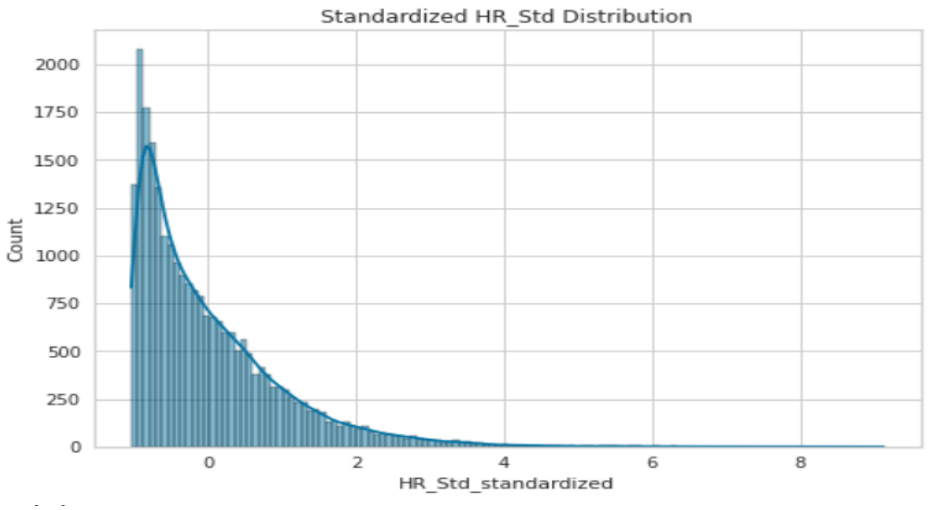
GNB is best suited for datasets with conditionally independent features, linear relationships, and normalized data (Bi et al., 2019). While it assumes conditional feature independence, which simplifies computation, this assumption can limit its performance with more complex, dependent features. The feature interactions that GNB can model are also linear, which is not the case for the complex relationship between many of these variables—for example, the interaction between the rolling sum of carbohydrates consumed and hours from midnight. Food can be consumed at the beginning of the day, which is less far from midnight, but that potentially increases CGM values in the subsequent windows of prediction.

KNN relies on distance measurements, which can be less effective with mixed data types and skewed distributions. Since most of these variables are skewed, KNN underperforms the tree models but outperforms GNB.

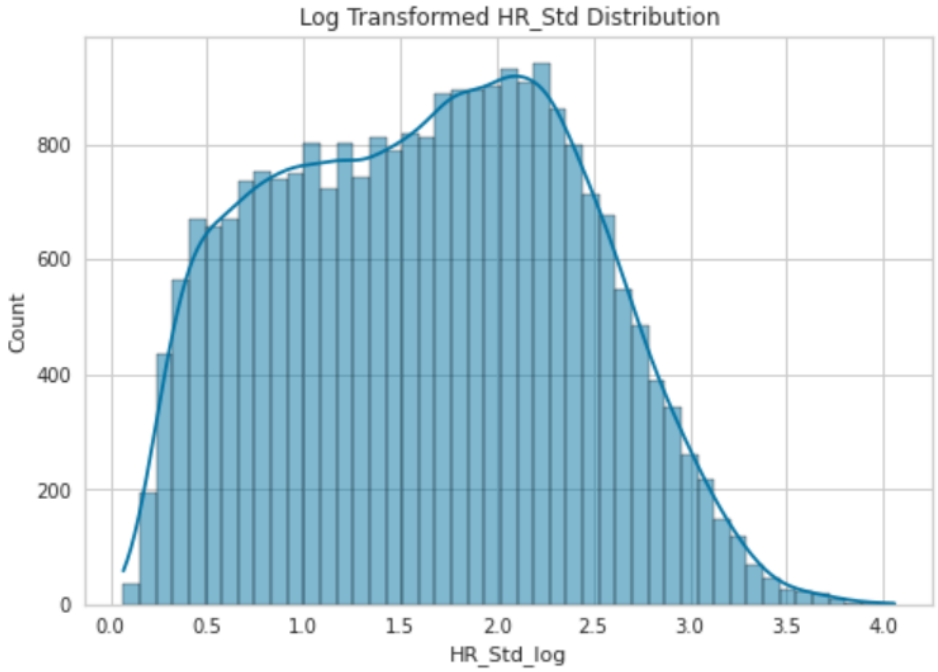
SVMs can capture complex feature relationships but often require extensive feature engineering to perform optimally (Cortes & Vapnik, 1995). For instance, standardization and transformations such as taking the logarithm of skewed features can improve SVM performance. Figure 4.13 illustrates the impact of standardization on the skewness of the heart rate (HR) standard deviation: normalization using the Z-score does not eliminate skewness, but applying a log transformation makes the changes more prominent. To illustrate the impact of skewness on model performance, models sensitive to skewness (SVM and GNB) are trained on the training set. The accuracy calculations on the validation set reveals that taking a log of HR_Std increases the accuracy of GNB (40% to 44%) and SVM (61% to 64%), although that increase is small.

Tree-based models are particularly adept at handling skewed features and mixed data types (Breiman, 2001). Of the tree models, ensemble methods like RF and XGBoost outperform other models by better handling outliers. The presence of influential outliers in the data is evidenced by the Cook's distance plot in Figure 15. RF's better performance over XGBoost in this study may be due to the noise in the smart watch data, suggesting that quality metrics should be used during data collection to minimize noise influence.

Similarly, tree-based models are better at learning mixed data types (categorical and numerical) than KNN models, because the latter rely on distance measurements. Amongst the three models, ensemble methods (RF and XGBOOST) performed better, because of their better handling of outliers.



(a)



(b)

Figure 4.13: Comparison of HR standard deviation skewness. (a) Normalization of HR values using the Z-score does not eliminate the skewness of the data. (b) Taking a log of this value makes the changes more prominent.

The preprocessing, feature engineering, and hyperparameter tuning in this work results in models with superior performance.

Tree-based models are effective in predicting glucose levels both for classification and regression. The explanations also provide a basis for future model development and feature engineering; for example, it is worthwhile to convert the skewed features to the log of these features or using PDP plots to engineer new features.

This study has potential limitations. The values of ground truth measured using a Dexcom sensor that define the labels are affected by motion (Laguna Sanz et al., 2019). However, the study clearly states the values it predicts.

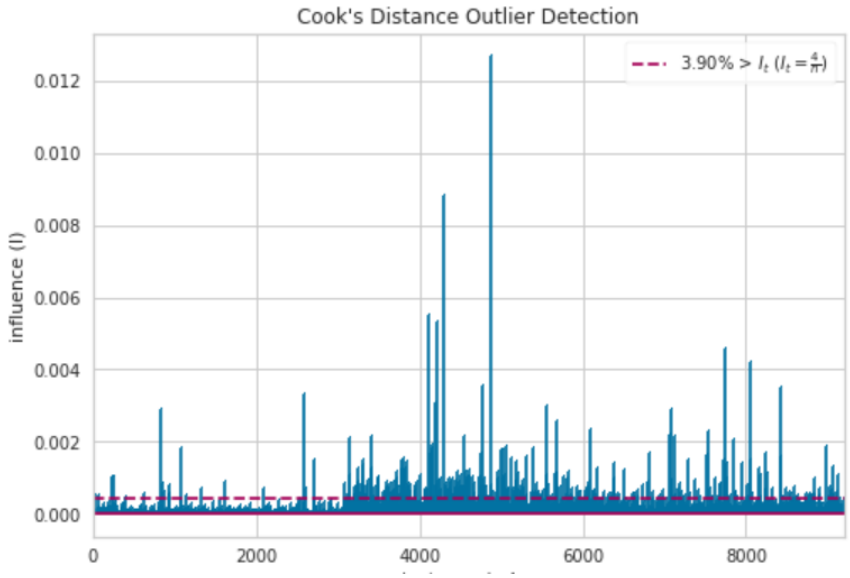


Figure 4.14: Cook's distance plot shows influential outliers.

Table 4.7 summarizes results of the best performing classification model (RF) in this work

Table 4.7: Performance of the RF model trained on tuned hyperparameters.

Class	Precision	Recall	F1-Score	Accuracy
High	0.80	0.85	0.82	0.80
Normal	0.71	0.89	0.79	0.70
Low	0.81	0.58	0.67	0.80

Table 4.8 compares the results of the best performing models in this work with earlier works, highlighting the superior performance of the models in this work.

Table 4.8: Comparison of the best performing models in this work with earlier works

Study	Type	Model	Performance
(Bent, Cho, Henriquez, et al., 2021)	Regression	DT	MSE = 21.22 ± 4.14 mg/dL
This Work	Regression	RF	MSE = 9.04 mg/dL
(Adams & Nsugbe, 2021)	Classification	DT	AUROC = 0.72
This work	Classification	RF	AUROC = 0.86

The models that are compared in this study are compared based on the features that have been engineered. Future feature engineering or postprocessing of the features, such as taking a log of the features, can affect the performance of different model types. These results underscore the importance of using robust ensemble methods for glucose level prediction, suggesting that these models can significantly improve the accuracy and reliability of real-time glucose-monitoring systems. In practical terms, the enhanced performance of these models can lead to better glucose management and improved health outcomes for individuals with diabetes. Additionally, this study's insights into feature engineering, such as the benefits of log transformation for skewed data, provide a valuable framework for developing more accurate predictive models in future research.

Implementing these findings in healthcare settings could facilitate more personalized and effective diabetes management, ultimately contributing to better patient care and quality of life.

4.7 Conclusion

This study has demonstrated that tree-based models, particularly RF and DT, exhibit superior performance in predicting IG levels from wrist-worn smart watch sensor data. These models outperformed other ML models in both classification and regression tasks, achieving higher accuracy, precision, recall, and F1-scores for classification, as well as lower RMSE and higher R-squared values for regression.

In conclusion, the findings of this study highlight the potential of using smart watch sensor data and tree-based ML models to provide insights into metabolic health and disease states. Future work should focus on improving data quality through noise reduction techniques and exploring advanced feature engineering methods, (e,g transforming skewed features). Implementing these improvements can further enhance the accuracy and reliability of IG level predictions, ultimately contributing to the better management of metabolic syndromes and diseases.

4.8 Chapter Summary

Chapter 4 answers research question three of this thesis. This research question is about identifying the best performing ML models for IG prediction using smart watch and food log data. The investigations presented in this chapter indicated that tree-based models are best suited for the regression and classification of IG levels using smart watch sensors for the given set of features that were identified in Chapter 3. Chapter 4 also explains the explanations of best performing models using SHAP values. SHAP explanations show the effect of hours from mid night and peak EDA values as the most important features for RF models. The importance of hours from midnight in predicting IG values also points to the effect of sleep activity cycles on IG values, that will be investigated in Chapter 5. Chapter 4 also examined the reasons for which tree models perform better at IG prediction from smart watch and food log data which include the ability of tree models to tackle outliers effectively and model complex relationships amongst features verified using PDPs and Cook's plots. This Chapter not just highlights the need for effective model development but also calls for better feature engineering and pre-processing techniques. The exploration in Chapter 5 will commence by designing novel sleep features that can be measured using smart watch sensors and their utility in developing better models.

5 Utility of Sleep Features in ML for Prediction of Interstitial Glucose

5.1 Preface

The content of this Chapter is a copy of the article “Utility of Sleep Features in ML for Prediction of Interstitial Glucose”, submitted and under review in the Journal of Computers in Biology.

In Chapter 5, new features are measured from smart watch sensors to increase the performance of interstitial glucose (IG) prediction models leveraging two distinct sensor combinations:

(a) Accelerometers to predict sleep parameters related to sleep duration,
(b) Heart Rate and Accelerometers based sleep stage prediction using a machine learning (ML) model.

After the broader literature review in Chapter 2, which highlighted how data can be converted into actionable insights, highlighting, preprocessing, feature or biomarker calculation, training a model, evaluating and explaining the model. Chapter 3 highlights the features of biomarkers measured by the time domain data measured using smart watches and food logs for predicting IG values using ML models. Chapter 2 and Chapter 3 led to two significant gaps in the literature; no works had compared the ML models for predicting IG values using empirical data and performance metrics and no work has used sleep related biomarkers or features in IG prediction models. This chapter addresses the second gap, by proposing different methods to calculate sleep features, measure their efficacy and compare the increase in performance in IG prediction statistically.

5.2 Abstract

This chapter demonstrates the use of sleep features estimated from wrist-worn accelerometers and heart rate sensors, to predict interstitial glucose (IG) levels using machine learning (ML). Previous studies have used activity, statistical, circadian and autonomic nervous system predictors from smart watch data in IG prediction ML models, but sleep features have not been utilized. Literature suggests a strong relationship between IG and sleep features. The novel contributions of this work are: 1) utilizing accelerometer data from smartwatches to find Number of Wake Bouts (#WB), Wake After Sleep Onset (WASO) , Sleep Onset Latency (SOL) and Total Sleep Time (TST) 2) utilizing accelerometer and HR data from smartwatches to train random forest model (RFSleep) to predict sleep stages Wake, Non-Rapid Eye Movement 1,2 and 3 and Rapid Eye Movement (W,N1, N2,N3,REM) based on a labelled public dataset 3) Transforming

sleep features to define additional features 4) Comparing the increase in performance of ML models (random forest (RF), decision tree (DT), support vector machine (SVM), Lasso, Ridge, and extreme gradient boosting (XGBoost)) for predicting glucose classes (classification) and values (regression). This paper compares the performance of models trained with and without sleep features using statistical tests (Paired T-tests, Friedman, and Nemenyi Post Hoc Analysis). The additional sleep features improve decision tree (DT) model performance by reducing mean absolute error (MAE) of predicted interstitial glucose (IG) from 8.0202 ± 0.2242 mg/dL to 6.5906 ± 0.3333 mg/dL (*p-value*: 0.0001). The accuracy of ML models in classification of IG values into high, low, and normal classes, increases from 0.7988 ± 0.0183 to 0.8265 ± 0.0111 (*p-value*: 0.0001). For the DT IG regression model, the average relative importance of sleep features is 12.8% and for DT IG classification model it is 14.5% underscoring the importance of sleep features in glucose prediction. These findings answer the research question four introduced in Chapter 1 which is about how sleep related parameters affect IG levels.

5.3 Introduction

Metabolic disorders such as prediabetes and diabetes are characterized by elevated glucose levels (Aguilar et al., 2015). Prediabetes is diagnosed when fasting glucose levels range from 100 to 125 mg/dL, glycated haemoglobin (HbA1c) levels are between 5.7% and 6.4%, or an oral glucose tolerance test yields results between 140 and 199 mg/dL (M. M. Kim et al., 2022). Diabetes on the other hand is defined by fasting glucose levels exceeding 126 mg/dL. Both diabetes and prediabetes are becoming increasingly prevalent; the World Health Organization (WHO) reported a rise in diabetes cases from 108 million in 1980 to 422 million in 2014 (Roglic, 2016). In the United States, 38% of the adult population is reported to have prediabetes (Zimmet et al., 2016). According to the Centre for Disease Control and Prevention (CDC), 81% of individuals with prediabetes are unaware of their condition (CDC, 2024), leading to high conversion rates from prediabetes to diabetes (Tabák et al., 2012a) Prediabetes is a reversible condition through lifestyle changes such as improved diet and increased physical activity (M. M. Kim et al., 2022). Therefore, monitoring glucose levels and their trends is crucial to help individuals reverse prediabetes and reduce the risk of progressing to diabetes.

Monitoring glucose levels is essential not only for those at risk of diabetes but also for individuals living with the condition. In diabetes, impaired regulation of blood glucose means that both low levels (hypoglycaemia , <70 mg/dL) and high levels (hyperglycaemia, >180 mg/dL) pose serious health risks (Patell et al., 2017). Traditionally, glucose values are measured through blood samples, particularly fasting HbA1c levels, which provide an overview of metabolic health over the past three to four

months (Bennett et al., 2007). However, regular blood sampling can be invasive and inconvenient (Jarvis et al., 2023). CGMs offer a less intrusive solution by measuring interstitial glucose (IG) levels every one to five minutes using a small filament inserted into the interstitial fluid (Poolsup et al., 2013). CGMs store data for up to eight hours and require regular downloads to perform analysis on data. CGMs have proven valuable for tracking the effects of sleep (Leproult & Van Cauter, 2010), physical activity (Colberg et al., 2010), dietary intake (Harvey et al., 2019), stress (Adam & Epel, 2007) and meal timing (Hutchison et al., 2019) on glucose variability and control. Therefore, monitoring glucose levels benefits not only individuals with diabetes but also those with prediabetes and even healthy individuals aiming to manage their glucose levels. However, CGMs need to be replaced every 14 days, and without integrating data on lifestyle factors like sleep, diet, stress, and activity, it is challenging to fully understand their impact on glycaemic control.

Given the limitations of CGMs—specifically their need for frequent replacement and lack of integration with lifestyle data—smartwatches like the Empatica E4 offer a non-invasive, continuously updating alternative for physiological monitoring. Equipped with sensors such as accelerometers, photoplethysmography (PPG), EDA, and temperature sensors, these devices can track metrics like physical activity, heart rate (HR), heart rate variability (HRV), sympathetic nervous system (SNS) activity, and thermoregulation. Unlike CGMs, smartwatches provide a constant stream of data without the need for regular maintenance and can capture lifestyle factors that impact glycaemic control. This comprehensive data collection makes smartwatches a promising tool for predicting glucose levels(Bent, Cho, Henriquez, et al., 2021).

Recent studies have demonstrated that smartwatch data, when combined with machine learning (ML) models, can effectively predict glucose fluctuations (Ali et al., 2023; Bent, Cho, Henriquez, et al., 2021; Jahromi et al., 2023; Patell et al., 2017; Zahedani et al., 2023). In these models, raw sensor data is converted into features such as mean values, frequency components, or dynamic time warping, which are more robust and informative for predicting glucose levels than the raw data itself (Breiman, 2001). These features are more closely related to the predicted value (label) than the raw sensor readings, which are prone to noise. In most of these studies, the labels come from CGM values. This approach not only allows for tracking glucose variability but also enables understanding it as a function of other physiological phenomena. A typical pipeline is shown in Figure 5.1.

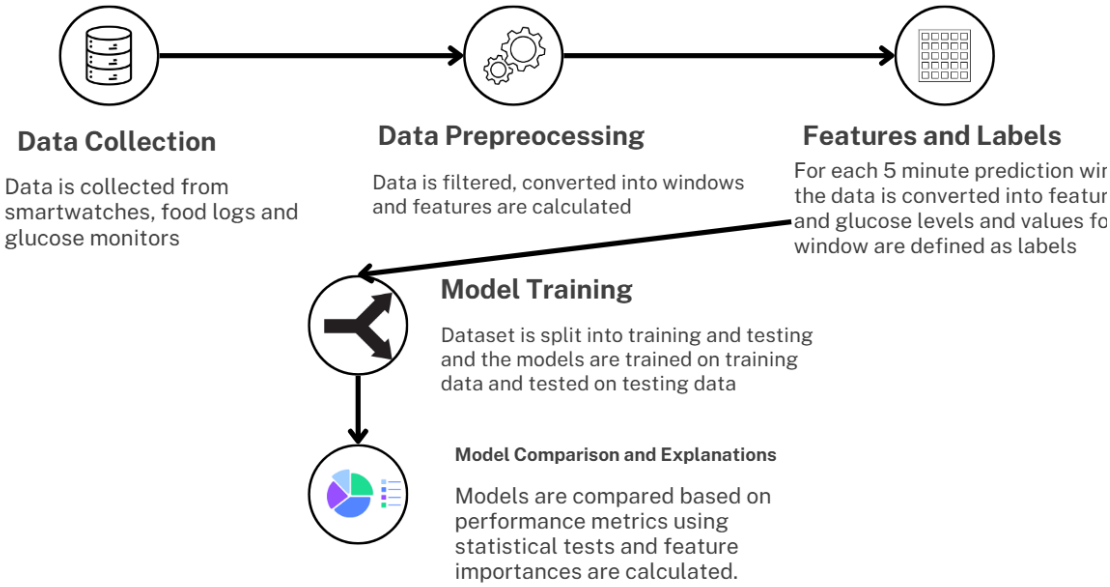


Figure 5.1: Stages of data processing in supervised learning for glucose prediction

To engineer features predictive of IG levels, raw sensor data values (e.g heart rate, inter-beat interval (IBI), skin temperature, and accelerometer readings collected from smart watch devices like the Empatica E4) are used. Thes raw signals are processed to extract key physiological predictors that correlate with glucose levels (e.g HRV, mean HR and activity intensity) (Bent, Cho, Henriquez, et al., 2021). Additionally, food log data is incorporated to calculate nutrient intake—including carbohydrates, fats, and proteins—which are known to influence IG levels. These features are then aggregated into specific time windows (e.g., 5-minute intervals) to align with CGM readings. Advanced preprocessing techniques like rolling averages and normalization are applied to smooth out the signals and highlight trends, making the features more suitable for machine learning models to predict glucose levels (Bent et al., 2021; Zahedani et al., 2023).

Sleep has also been shown to affect IG levels. According to (Knutson et al., 2007) and (Leproult & Van Cauter, 2010) sleep influences glucose levels through three physiological mechanisms: alteration of glucose metabolism, upregulation of appetite, and reduced energy expenditure. Sleep is divided into 30-second epochs, each categorized into one of three basic stages: rapid eye movement (REM), non-rapid eye movement (NREM), and wakefulness. NREM sleep is further divided into two light stages (N1, N2) and a deep sleep stage (N3) (Cole et al., 1992). N1 acts as a transitional phase, while N2 constitutes most of the sleep period and plays a key role in memory consolidation. N3, also known as slow-wave sleep (SWS), is the most restorative phase, crucial for physical recovery and immune function. REM sleep, characterized by vivid dreaming and heightened brain activity, is essential for cognitive functions like learning

and emotional regulation. Together, these stages form the architecture of a sleep cycle, with REM becoming more prevalent as the night progresses.

A substantial body of literature investigates the effect of each sleep stage on glucose metabolism (Briançon-Marjollet et al., 2015; Herzog et al., 2013). SWS is particularly important, as it is associated with reduced brain glucose utilization, increased parasympathetic activity, and elevated growth hormone levels—all of which help maintain stable glucose levels during the night. Studies have shown that insufficient SWS, due to sleep deprivation or fragmentation, impairs glucose tolerance and insulin sensitivity, increasing the risk of type 2 diabetes (Leproult & Van Cauter, 2010). The balance between REM, NREM, and wakefulness also affects glucose utilization, with metabolism slowing during SWS and increasing during REM sleep and wakefulness.

Sleep-related metrics are typically measured using polysomnography (PSG), a comprehensive sleep study used to diagnose sleep disorders by recording various physiological parameters during sleep (Chase et al., 2022). PSG involves monitoring brain activity via electroencephalography (EEG), eye movements via electrooculography (EOG), muscle activity via electromyography (EMG), heart rate, respiratory effort, airflow, oxygen levels in the blood, and body movements. The data collected provides detailed information on sleep stages, breathing patterns, and the presence of abnormalities such as sleep apnoea, restless leg syndrome, or insomnia.

While PSG offers detailed insights, it requires clinical settings. In contrast actigraphy which relies on wrist worn accelerometer values can be used in daily life for monitoring human rest and activity cycles(Coyle-Asbil et al., 2023; Migueles et al., 2019). Actigraphy is commonly used to assess sleep patterns—including sleep duration, quality, and disturbances day by comparing their rates of change in accelerometer values which record wrist movement. Rule based methods based on thresholds of movement levels, such as those implemented in the GGIR R package (Migueles et al., 2019) are developed to measure sleep parameters from accelerometer data. Additionally, PSG-labelled datasets like the one provided by (Walch, 2019) can be used to train ML models that can discern sleep stages each epoch from HR and accelerometer data.

Recent works have utilized autonomic nervous system features (ANS) features, statistical features, food features, circadian features and activity features (Adams & Nsugbe, 2021; Ali et al., 2023; Bent, Cho, Henriquez, et al., 2021). But no works to our knowledge have used smart watch data to measure sleep features predictive of glucose levels.

This work thus adds the following novel contributions:

1. Conversion of accelerometer and heart rate data into sleep features (wake bouts (#WB), wake after sleep onset (WASO), sleep onset time (SOT), total sleep time (TST)) predictive of interstitial glucose (IG) classes.

2. Training an ML model (RFSleep) to predict sleep stages (W, N1, N2, N3, REM) using a public dataset and applying RFSleep to determine sleep stages in another dataset.

3. Enhancing sleep and time related features to make them more informative.

4. Comparing the performance improvement of various IG-predicting ML models with the inclusion of sleep features.

5.3.1 Related work

As highlighted in the previous section, recent studies have utilized features derived from ANS, statistical, food , circadian, and physical activity features in ML models to predict IG levels (Bent, Cho, Wittmann, et al., 2021b). However, none have incorporated sleep features derived from smartwatch data. This section reviews existing works that employ ML models to predict IG values and classes using smart watch sensor data. Several studies have explored the use of ML for IG level prediction:

- (Bent, Cho, Henriquez, et al., 2021)used the Empatica E4 smartwatch, food logs, and Dexcom CGMs to collect inter-beat interval (IBI), HR , blood volume pulse (BVP), EDA, temperature, and food intake data. They engineered time, ANS, statistical, and food features to train a decision tree (DT) model, achieving a root mean squared error (RMSE) of 21.22 ± 4.14 mg/dL in predicting IG values.

- (Ali et al., 2023) also utilized the Empatica E4, food logs, and Dexcom CGMs to collect similar sensor data. They extracted time, ANS, statistical, and food features to train support vector machine (SVM) and DT models for IG levels classification, achieving accuracies of $68.96 \pm 14.4\%$ and $72.38 \pm 2.4\%$, respectively.

- (Zahedani et al., 2023) employed Apple Watch or Fitbit devices, food logs, and FreeStyle Libre CGMs to collect HR, activity, time, ANS and food intake data. They used a long short-term memory (LSTM) layer to convert timestamped data into features and trained an LSTM model to predict IG values, achieving an RMSE of 14.8 mg/dL.

- (Ali et al., 2024) continued this line of research using the Empatica E4, food logs, and Dexcom CGMs. They engineered time, ANS, statistical, and food features to train an extreme gradient boosting (XGBoost) model, achieving individual-specific accuracies ranging from 64% to 86%.

- (Zhu et al., 2022) used the same devices and sensor data as previous studies, extracting ANS, statistical, and food features to train a bidirectional recurrent neural network (BRNN), achieving an RMSE of 35.28 ± 5.77 mg/dL.

These studies generally define glucose classes by transforming IG values into personalized high, normal, and low categories. Specifically, IG values within one standard deviation of the average are labelled as normal, values above one standard deviation as high, and those below as low (Bent, Cho, Henriquez, et al., 2021). This is shown in Figure 5.2

Table 5.1 compares different ML pipelines used in predicting IG values and classes from smart watch data.

While these studies have made significant contributions, they primarily focus on physiological, and lifestyle features excluding sleep metrics. Other works aim to predict hypoglycaemia and hyperglycaemia using similar sensor inputs for diabetic populations (Abbas et al., 2018; Bertachi et al., 2020; Lee et al., 2023; Patell et al., 2017).

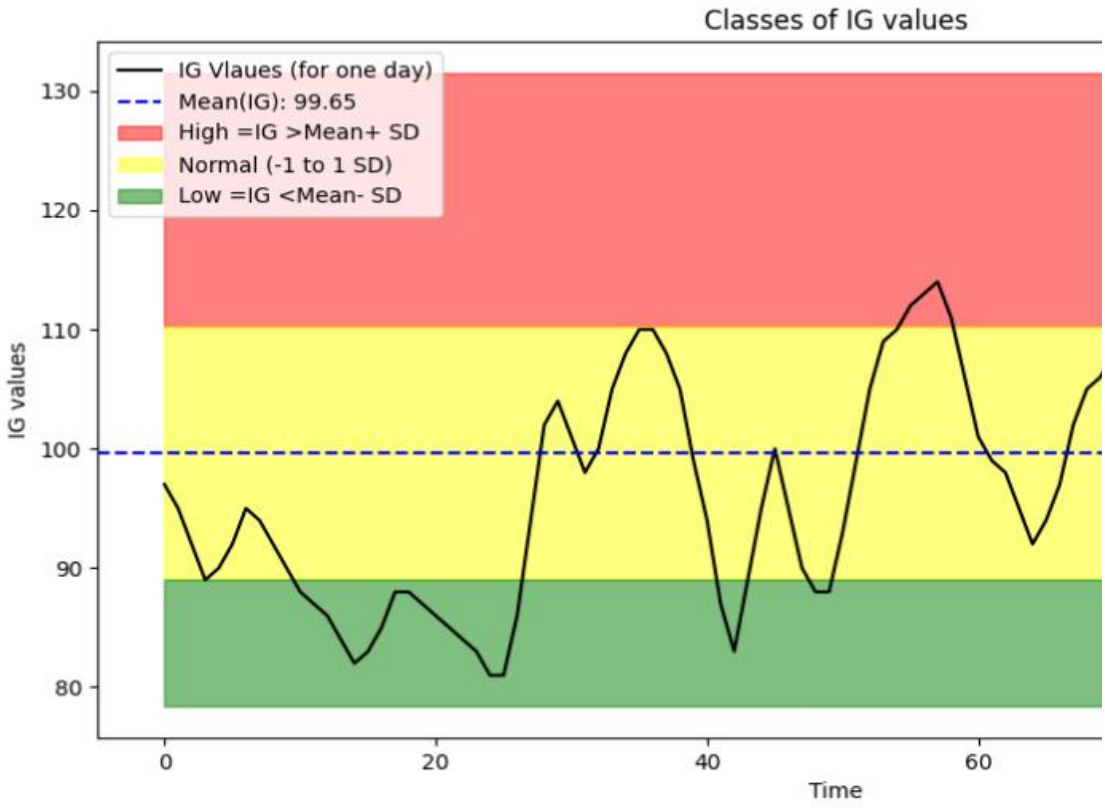


Figure 5.2: Definition of classes of IG values

Despite the advancements, none of the aforementioned studies have integrated sleep features calculated from accelerometer or heart rate data into their ML models. Additionally, there is a need to modify time-related features to ensure continuity and improve prediction accuracy.

Table 5.1: Comparison of ML models for prediction of glucose levels from smart watch**es**

Work	Devices	Sensors	Feature Types	Models	Performance Metrics
(Zhu et al., 2022)	Empatica E4, Food Log, and Dexcom CGM	Inter-beat interval (IBI), Heart Rate (HR), Blood Volume Pulse (BVP), EDA, Temperature (T), Food Log.	ANS, Statistical, and Food Features	Bidirectional Recurrent Neural Network (RNN)	RMSE = 35.28 ± 5.77 mg/dL
(Bent, Cho, Henriquez, et al., 2021)	Empatica E4, Food Log, and Dexcom CGM	Inter-beat interval (IBI), Heart Rate (HR), Blood Volume Pulse (BVP), EDA, Temperature (T), Food Log.	Time, ANS, Statistical, and Food Features	Decision Tree (DT)	Root Mean Squared Error (RMSE= 21.22 ± 4.14 mg/dL)
(Ali et al., 2023)	Empatica E4, Food Log, and Dexcom CGM	Inter-beat interval (IBI), Heart Rate (HR), Blood Volume Pulse (BVP), EDA, Temperature (T), Food Log.	Time, ANS, Statistical, and Food Features	Support Vector Machine (SVM), DT	SVM (Accuracy = 68.96 ± 14.4 %), DT (Accuracy = 72.38 ± 2.4 %).
(Zahedani et al., 2023)	Apple Watch or Fitbit, Food Log and FreeStyle Libre CGM	HR and Food Log.	Long short-term layer (LSTM) converts time stamped data into features	LSTM	RMSE= 14.8 mg/dL
(Ali et al., 2024)	Empatica E4, Food Log, and Dexcom CGM	Inter-beat interval (IBI), Heart Rate (HR), Blood Volume Pulse (BVP), EDA, Temperature (T), Food Log.	Time, ANS, Statistical, and Food Features	Extreme Gradient Boosting (XGBoost)	Accuracy= 64% to 86% (individual specific)

This research seeks to address these gaps by integrating sleep features, calculated from accelerometer and HR data, into ML models for predicting IG levels in healthy participants. By transforming smart watch sensor data into meaningful sleep metrics and training ML models that incorporate these features, this work aims to demonstrate the added value of sleep data in improving IG prediction accuracy. This research divides sleep features into two classes: sleep parameters and sleep stages. Sleep parameters include wake bouts (#WB), wake after sleep onset (WASO), sleep onset time (SOT), total sleep time (TST)), these are measured using accelerometer values using rule-based methods and sleep stages: W, N1, N2, N3 and REM which are estimated with the help of ML model (RFSleep). RFSleep is trained using a PSG labelled public dataset.

5.3.2 Datasets

To address the gaps identified in previous research and to demonstrate the utility of sleep features in predicting IG levels, we utilize two public datasets. The first dataset is used to train RFSleep model for sleep stage prediction based on smart watch sensor data, and the second dataset is employed to evaluate the effectiveness of incorporating sleep features into IG prediction models.

5.3.2.1 Labelled Sleep Dataset

To train RFSleep to predict sleep stages, motion and HR data from a wrist-worn smart from a PSG labelled dataset is used (Walch, 2019). The dataset includes 31 participants. Participants, recruited from the University of Michigan, wore the Apple Watch for 7-14 days. The dataset includes timestamped files for each participant: acceleration (in g), HR (in bpm), Steps Data, and PSG labels for each epoch. Each line in PSG file has the format: date (in seconds since PSG start), stage (0-5, wake = 0, W, N1 = 1, N2 = 2, N3 = 3, REM = 5).

5.3.2.2 Glucose Prediction Dataset

To evaluate the effectiveness of incorporating sleep features into IG prediction models, we use a public dataset provided by(Cho et al., 2023a). This dataset has 16 participants, aged 35-65, who have elevated blood glucose levels within the normal to prediabetic range. The cohort includes 9 females and 7 males. Participants wore a Dexcom G6 CGM and an Empatica E4 wristband for 8-10 days to capture smart watch data. Standardized breakfast meals were provided every other day, and participants maintained a food log. Data was date-shifted to ensure anonymity. The dataset includes timestamped files for each participant:

- The ACC file records accelerometer data for the X, Y, and Z axes.

- The BVP file contains blood volume pulse measurements.
- Dexcom, EDA, TEMP, IBI, and HR files provide glucose values, electrodermal activity, skin temperature, inter-beat intervals, and heart rate data, respectively.
- The Food Log file details the consumed food items, including date, time, type, amount, calories, carbohydrates, dietary fiber, sugar, protein, and fat content.

Demographic data, including gender and HbA1C values, are also included. PPG data is sampled at 64 Hz for heart rate and blood volume pulse, EDA, and skin temperature at 4 Hz, and accelerometry at 32 Hz. The CGM records glucose classes every 5 minutes. This dataset allows us to apply the sleep stage predictions from the RFSleep model to real-world physiological data and assess the impact of sleep features on the performance of IG prediction models.

5.3.3 Comparison of Methods for sleep parameter estimation

To ensure repeatability in estimating sleep parameters, this study employs rule-based methods that utilize accelerometer data from wrist-worn devices. Several methods have been proposed in the literature for deriving sleep metrics from accelerometer readings. For examples, Liu et al., 2020 uses disaggregation of steps data by assuming a uniform distribution of steps during sleep to find sleep onset and (Grandner et al., 2023) uses the steps and heart rate from Fitbit Alta to estimate sleep parameters but these are not selected as they use step values and not accelerometer values. (Sundararajan & Hees, 2020) use ML to identify sleep parameters from accelerometer data but does provide the model weights and are therefore not suitable for our purposes.

In this work, we have selected rule-based methods that rely solely on accelerometer data to estimate sleep parameters. Table 5.2 shows a summary of these works and how each sleep parameters effects glucose level.

Table 5.2: Rule based methods to identify sleep features from wrist worn accelerometer data

Feature	Definition	Implementation
Wake after sleep onset (WASO)	WASO represents the total duration of wakefulness occurring after the initial onset of sleep. It serves as an indicator of sleep fragmentation, with higher values suggesting more frequent awakenings and potentially poorer sleep quality. Elevated WASO is commonly associated with sleep disorders such as insomnia.	(Plekhanova et al., 2023) and GGIR package in R can be used to measure WASO based on accelerometer data
Total Sleep Time (TST)	TST denotes the cumulative amount of actual sleep time during a designated sleep period, excluding any periods of wakefulness. It is a critical measure of sleep quantity, with adequate TST being essential for maintaining optimal physical and mental health. Insufficient TST can contribute to a variety of adverse health outcomes.	(van Hees et al., 2018) uses thresholds on motion of the wrist to determine the wake and sleep times and the total sleep time.
Percent Sleep Time (PST)	PST, calculated as (Total Sleep Time / Time in Bed) * 100%, indicates the proportion of time spent in bed that is spent sleeping. This measure reflects sleep efficiency, with higher percentages signifying a more efficient sleep pattern where a greater proportion of time in bed is utilized for sleep.	(van Hees et al., 2018) and (Migueles et al., 2019) is used to measure the percent sleep time.
Number of wake bouts (#WB)	#WB refers to the total count of awakenings occurring during the sleep period. An increased number of wake bouts can signify disrupted sleep, as frequent awakenings can interfere with the continuity of the sleep cycle and diminish overall sleep quality.	(Bai et al., 2016) is used. It uses heuristic algorithm based on arm movement to determine the presence of wakefulness in sleep using accelerometer values.
Sleep Onset Latency (SOL)	SOL measures the time taken to transition from wakefulness to sleep, typically recorded from the moment an individual attempts to fall asleep to the actual onset of sleep. Prolonged SOL may indicate difficulties in initiating sleep, a common symptom in conditions such as insomnia. Conversely, a short SOL may suggest excessive sleepiness.	(Migueles et al., 2021) describes the estimation of sleep onset latency based on (Migueles et al., 2019)

5.4 Materials and Methods

In this study, we employ rule-based methods to convert accelerometer data from(Cho et al., 2023a) into sleep parameters. RFSleep model is used to predict sleep stages for each night. We calculate an additional feature set following previous studies (Ali et al., 2023; Bent, Cho, Henriquez, et al., 2021) and incorporate the calculated sleep features into this set. Performance of IG predicting ML models are compared with and without sleep features. The following sections outline each step from data preprocessing to performance evaluation.

5.4.1 Preprocessing Steps

For the two datasets used in this work, we apply preprocessing methods recommended in the literature. Accelerometer data is filtered using a Butterworth low-pass filter with a cutoff frequency of 30 Hz (Jahromi et al., 2023). Heart rate data is filtered with a bandpass filter within the range 0.5-4Hz (Brunner & Hofer, 2023). For IBI, BVP, EDA and T data, Savitzky-Golay filter, and weighted moving average filter are used (Aryal & Becerik-Gerber, 2019; Mbarek et al., 2022). For sleep features, the heart rate and accelerometer data must be converted into 30 second windows(epochs), whereas for calculating features for glucose prediction, 5-minute windows are used. This filtering and windowing were done in python using NumPy (Harris et al., 2020), Pandas (McKinney & Team, 2015) and Seaborn.

5.4.2 Implementation of sleep features

WASO, SOT, TST, PST and #WB are calculated based solely on accelerometer data. Sleep windows are detected based on (Hees et al., 2015; Migueles et al., 2019) . Sleep and wake detections are based on (Cole et al., 1992) whereas activity indices are measured using (Bai et al., 2016). From the sleep and wake detection WASO, SOT, TST, PST and #WB are calculated. This is implemented using SleepPy public application interface (API) (Christakis, 2019/2024). SleepPy is designed to be implemented for GeneActiv data, but the glucose labelled dataset (Cho et al., 2023b) has Empatica E4 data. To make the Empatica E4 data suitable for SleepPy operation, the following modifications are made: time column is changed from Portable Operating System Interface (POSIX) to Coordinated Universal Time (UTC), adding in *Lux button* with all zeroes, so columns match up with GeneActiv although Lux is not used in the code, dividing x and y by 64 to convert to gs (*2/128 as Empatica E4 records in units of 1/64 Gs). The dates for which the data was available for the night, WASO, SOT, TST, PST and #WB were calculated. After making the modifications to E4 data, SleepPy is used to measure the sleep parameters (WASO, SOT, TST, PST and #WB).

For detecting the sleep stages in the night, RFSleep was trained to predict the stage of sleep per epoch (30 second window). The model was trained using the pipeline described (Walch et al., 2019). In this pipeline three kinds of features are used: Motion features, HR features and Time features. The data (accelerometer and HR) was converted into the respective features alongside the time features. The data used for training RFSleep is from apple watch (HR and acceleromter) and is labelled using PSG (Walch, 2019). The accelerometer data from (Walch, 2019) is in the units of fractions of g. After training and validating RFSleep using (Walch, 2019), the trained RFSleep is used to infer sleep stages from (Cho et al., 2023). (Cho et al., 2023) unlike (Walch, 2019) contains Empatica E4 data. The differences between these two datasets (Cho et al., 2023 and Walch, 2019) are resolved using the following strategy. Firstly, data availability from (Cho et al., 2023) for the nights is identified, after which, sleep windows are detected using (Hees et al., 2015). (Walch, 2019) contains time stamps in the form of difference from sleep labelling time (sec), negative for times before PSG labelling and positive for after that time. The timestamps of the data from (Cho et al., 2023) are transformed such that they are in the form of difference in seconds from sleep onset detected using (Hees et al., 2015). The values of accelerometers are transformed so that they are in the form of fractions of g. For each 30 second window (sleep epoch), this transformed data of (Cho et al., 2023) was converted into HR, Motion and Time features described in (Walch et al., 2019). RFSleep model was used to predict sleep stages per epoch.

After finding sleep parameters (WASO, SOT, TST, PST, #WB) and sleep stages (W, N1, N2, N3, REM) for the nights in (Cho et al., 2023a). Sleep features are input alongside *other features* from (Ali et al., 2024; Bent, Cho, Henriquez, et al., 2021) in all the five-minute windows of the next day. These sleep features are used to define new features per night. The number of epochs for each stage per night are counted (W_count,N1_count, N2_count, N3_count, REM_count). Each sleep stage is converted into a percentage value of the total sleep time (W_percent, N1_percent, N2_percent, N3_percent, REM_percent). The difference from the recommended sleep percentage per stage is calculated (W_difference_from_recommended, N1_difference_from_recommended, N2_difference_from_recommended, N3_difference_from_recommended, REM_difference_from_recommended). The skewness of the new features is tested, and a correlation of sleep feature is measured with the IG value. The sleep features are further transformed using different statistical techniques (Yeo, Logarithm etc.) and the corresponding correlations with the IG values are measured. The transformations that result in transformed features with highest correlation with IG values are selected.

To enhance the time features in the form of time stamps in the dataset, the timestamps are converted into its constituents (Day, Hour, Minutes and Seconds). A sine and cosine for all of constituents of time stamps are also calculated as features predictive of interstitial glucose.

Thus there are three kinds of features named other features defined in (Ali et al., 2024; Bent, Cho, Henriquez, et al., 2021), unmodified sleep features and transformed sleep features. The feature importances and shapely additive explanations of DT model are used to qualify the importance of each feature type.

5.4.3 ML Models for IG Predictions

The following models are used for predicting IG levels (classes and values). The first model used for that is Random Forest (RF). RF is an ensemble learning method that trains multiple DT during training and outputs the average prediction of the individual trees for regression tasks, or the mode of the classes for classification tasks. Each tree is trained on a different subset of the data, created by sampling with replacement. At each split in a tree, a random subset of features is considered. RF is robust to overfitting, handles large datasets well, and works well with both numerical and categorical features (Breiman, 2001). The second model used for IG level prediction is DT. DT is a non-parametric model used for classification and regression. It splits the data into subsets based on the value of input features, creating a tree-like structure of decisions. DT comprises nodes (decisions), branches (outcomes), and leaves (final output). Decisions are based on Gini impurity or entropy for classification and mean squared error for regression (Kingsford & Salzberg, 2008). The next model used for IG level prediction for each window is SVM. SVM is a supervised learning model that analyses data for classification and regression analysis by finding the hyperplane that best separates the data into classes. Hyperplane is the decision boundary that maximizes the margin between different classes. With the help of the Kernel trick SVM transforms data into a higher dimension where it is easier to find a separating hyperplane. SVMs are effective in high-dimensional spaces, robust to overfitting, especially in high-dimensional space. However, SVMs are memory-intensive, and the choice of kernel and regularization parameters requires careful tuning (Park et al., 2022). The next model used is called Lasso Regression. Lasso (Least Absolute Shrinkage and Selection Operator) is a linear regression method that includes an L1 penalty, encouraging sparsity in the model coefficients. L1 penalty equals to the absolute value of the magnitude of coefficients. Lasso Automatically selects important features by shrinking less important feature's coefficients to zero. It can handle large numbers of features, useful for feature selection. Lasso can struggle with highly correlated features and may not perform as well as Ridge when all predictors are relevant (Ranstam & Cook, 2018). Another method used for IG

level prediction is Ridge Regression. Ridge regression is a type of linear regression that includes an L2 penalty, which adds the squared magnitude of coefficients as a penalty term. L2 Penalty is a penalty equal to the squared value of the magnitude of coefficients. This helps prevent overfitting by shrinking coefficients. Ridge models can handle multicollinearity, useful when there are many predictors. Ridge does not perform feature selection as coefficients are only shrunk, not eliminated (Nouretdinov et al., 2001).The last model used for predicting IG values is extreme gradient boosting (XGBoost). XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. Gradient Boosting sequentially builds trees where each new tree corrects errors made by the previous ones. XGBoost includes both L1 (lasso) and L2 (ridge) regularization to improve model generalization. XGBoost High performance, ability to handle missing values, supports parallel and distributed computing, highly customizable. However, XGBoost Can be prone to overfitting if not properly tuned, and can be computationally intensive (T. Chen & Guestrin, 2016).

The regression models are compared based on the following performance metrics, MAE (Mean Absolute Error) quantifies the average absolute differences between the predicted and observed IG values, quantifying the error in predicted values. Similarly, RMSE (Root Mean Squared Error) represents the square root of the mean of the squared differences between predicted and observed IG values. MAPE (Mean Absolute Percentage Error) calculates the average percentage difference between the actual IG values and the predicted values. R^2 (R-squared) and Adjusted R^2 are performance metrics that assess the proportion of variance in observed IG values that is explained by the input features , with higher values indicating a better model fit. MSLE (Mean Squared Logarithmic Error) measures the average squared logarithmic differences between the actual and predicted IG values, making it less sensitive to outliers.

For classification tasks, first the data is stratified to ensure that number of samples are the same per class (high, low and normal) (2500x3=7500 samples) and the following metrics evaluate the models' effectiveness: Accuracy shows the proportion of correctly predicted instances out of the total number of instances. Precision is the ratio of true positive (TP) predictions to the total number of actual positive predictions (False Positive (FP) +TP). Recall measures the proportion of actual positives that are correctly identified by the model, calculated as TP / (TP + FN). The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. ROC (Receiver Operating Characteristic) plots the true positive rate against the false positive rate at various thresholds, with the AUC (Area Under the Curve) indicating the model's ability to distinguish between classes, where a higher AUC signifies better performance.

To determine if the performance metric differences are statistically significant, the Friedman test is used. The dataset is divided into 10 folds; the model is trained on 9 folds and tested on the remaining fold, iteratively, each time validating on a new set and training on other sets. The performance metrics are then analysed with the Nemenyi post hoc test to interpret the differences. The performance metrics for each fold are used to determine the performance. For comparison of model performance for different model types, and feature combinations (*other features* are features defined in Chapter 3) (other features, other features+ unmodified sleep features and other features + modified sleep features), statistical tests are used as described in (Rainio et al., 2024). For comparison of different models based on performance metrics Friedman test and Nemenyi Post Hoc Analysis is used. For comparison of different feature sets, paired t-tests of performance metrics for a DT model are used.

5.5 Results

Figure 5.3 shows the participants for which the night data is available based on the time stamps. The night data is not available for Participant 001.

The sleep parameters (WASO, SOT, TST, PST and #WB) are calculated with the help of SleepPy (Christakis, 2019/2024). The accelerometer, time and temperature values are modified to ensure compatibility with SleepPy library. These sleep parameters are added as features in all the 5-minute prediction windows for the following day

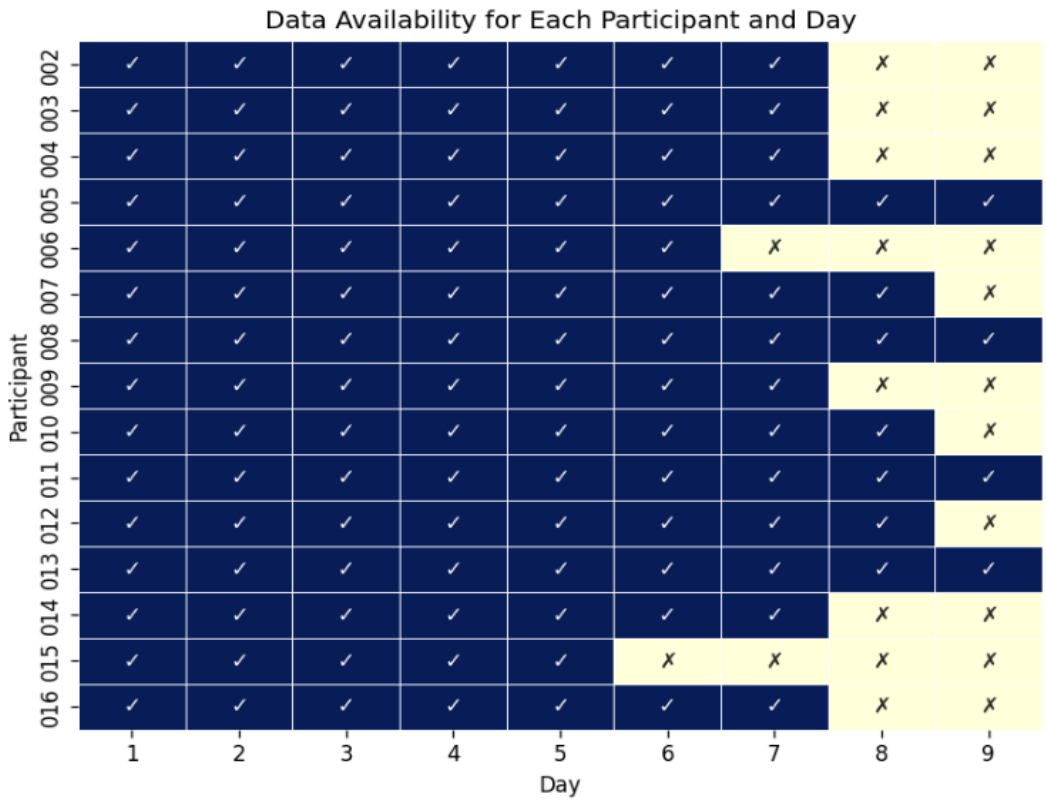


Figure 5.3: Availability of sleep data for different participants in (Cho et al., 2023a)

Table 5.3 and Figure 5.4 shows the statistical properties of the sleep parameters measured. The missing sleep parameters are imputed with Multiple Imputations with Chained Equations (MICE) (Z. Zhang, 2016) using only accelerometer values.

Table 5.3: Statistical features for available sleep data

Statistic	Mean	Standard Deviation	Maximum	Minimum
TST	420 mins	60 minutes	528 minutes	260 Min
PST	85%	5%	92%	68%
WASO	45 minutes	15 minutes	10 minutes	55 minutes
SOT	16.67 minutes	8.67 minutes	30.4 minutes	4.87 minutes
#WB	20	10	40	5

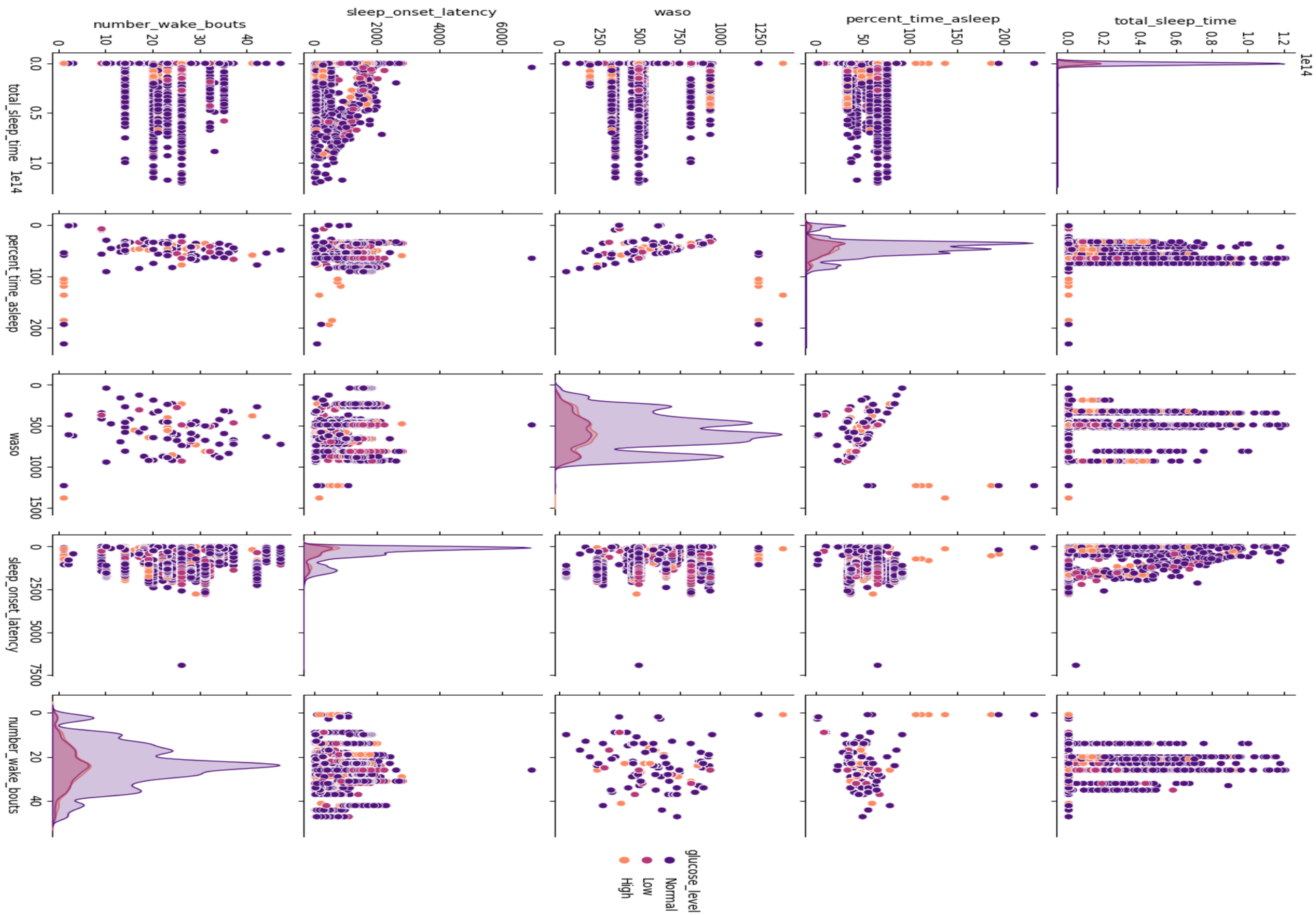


Figure 5.4: Pair plot of sleep parameters

5.5.1 Sleep Stage Classification with RFSleep Model

The RFSleep model for sleep stages is trained on the (Walch et al., 2019) dataset. Features of the PSG labelled data are calculated using the pipeline described (Walch et al., 2019). This dataset has a class imbalance as visible in the Figure 5.5.

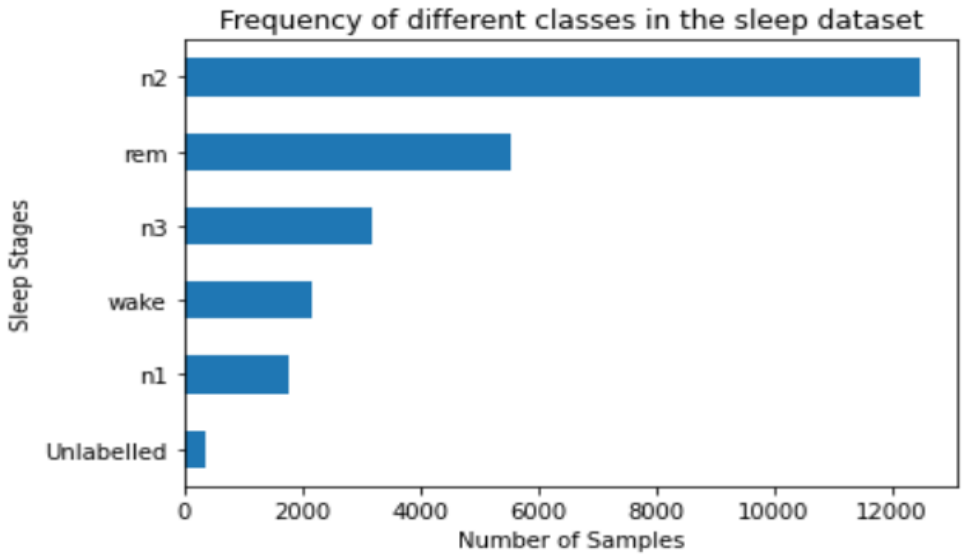


Figure 5.5: Class Imbalance in sleep dataset

We randomly under sampled each class to 2,000 samples. 70% of data is used to train the RFSleep model, whereas 30% data is used to test it. Model performance after balancing the dataset and standardization for classification of sleep stages shown in Figure 5.6 and Figure 5.7.

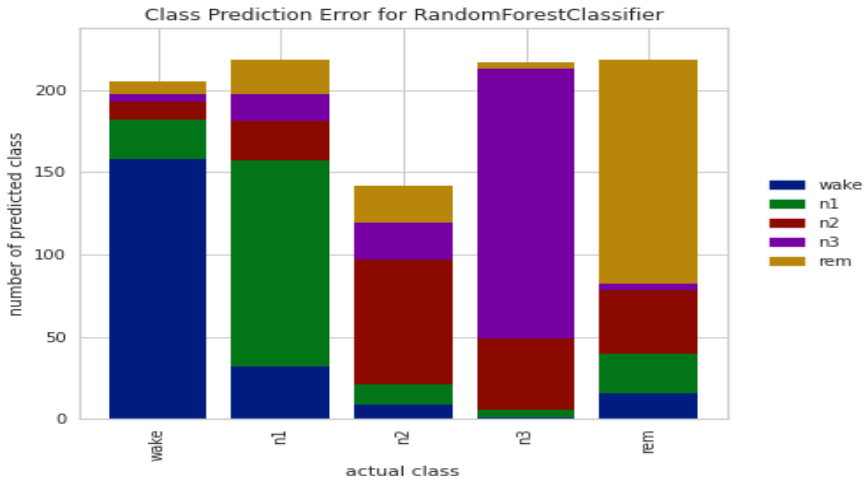
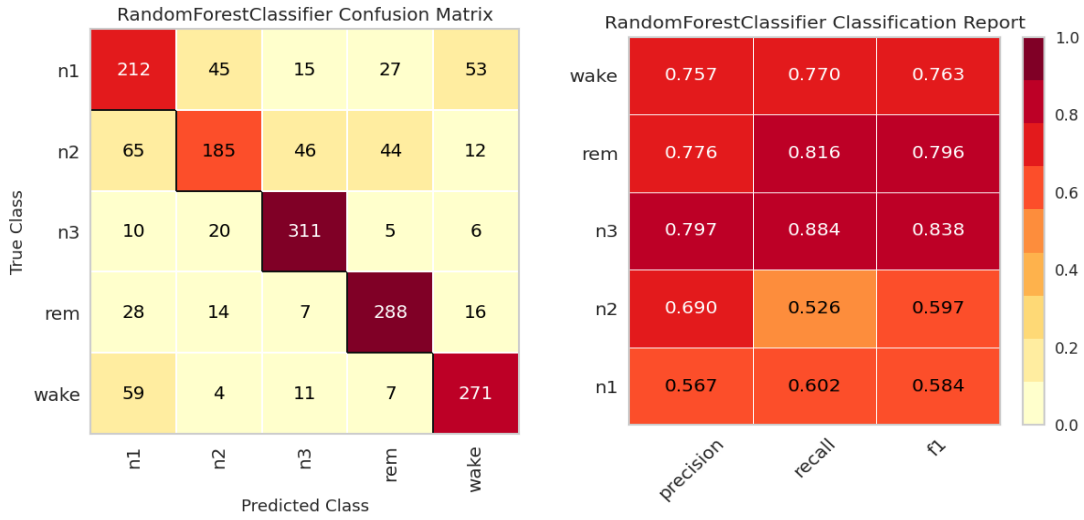


Figure 5.6: Class Prediction Error for RFSleep in sleep stage classification

The hyperparameters of the model are trained using Bayesian Optimization with the help of Optuna (Akiba et al., 2019) The optimized hyperparameters of this model are given in Table 5.4

Table 5.4: Optimized Parameters of RFSleep model

Hyperparameter	Number of Estimators	Maximum Depth	Minimum Sample Split	Minimum Leaves Per Sample
Value	130	22	7	2



(a) (b)

Figure 5.7: Performance of RFSleep model. (a) Confusion Matrix of sleep stages (b) Performance metrics (precision, recall, and f1-score) of RFSleep for each sleep stage

5.5.2 Sleep Stage Inference on Glucose Prediction Dataset

To apply the RFSleep model to the glucose prediction dataset (Cho et al., 2023a) the we adjusted the accelerometer and HR data to match the sampling frequency of (Walch et al., 2019). Inference was performed for all nights with available sleep data, as indicated in Figure 5.4. Table 5.5 shows the statistical properties of the sleep stages predicted by RFsleep on (Cho et al., 2023a) .

Table 5.5: Statistical properties of sleep stages measured from available sleep data from (Cho et al., 2023a)

Statistic	Mean	Standard Deviation	Maximum	Minimum
Wake	12.4	5.2	58	2
REM	43.7	7.2	60	5

N1	12.8	8.1	28	3
N2	108.7	11.43	123	80
N3	38.1	6.2	45	5

These parameters are used to define further sleep features

5.5.3 Modified sleep features

To enhance the predictive power of the sleep features, we conducted two statistical analyses:

1. Normality Assessment: We performed the Shapiro-Wilk test to determine whether each sleep feature followed a normal distribution. This informed our choice between using Pearson's or Spearman's correlation coefficients when examining relationships with interstitial glucose (IG) values.

Table 5.6: Shapiro Wilk test for checking the normality of sleep features

Feature	Statistic	*P* Value
TST	0.166	4.63×10^{-133}
PST	0.21	2.55×10^{-131}
WASO	0.98	1.27×10^{-45}
SOL	0.68	1.35×10^{-106}
#WB	0.98	8.44×10^{-38}
Wake	0.21	3.66×10^{-131}
N1-Epochs	0.21	3.66×10^{-131}
N2-Epochs	0.21	3.66×10^{-131}
N3-Epochs	0.21	3.66×10^{-131}
REM-Epochs	0.21	3.66×10^{-131}

2. Skewness Evaluation: We calculated the skewness of each sleep feature to understand the asymmetry of their distributions, which guided potential data transformations to improve normality and correlation strength.

To begin the analysis, the results of Shapiro-Wilk test are given in Table 5.6

The values of the statistic farther from 1 (p-values all less than 0.05) indicate that only WASO and #WB are normally distributed. Therefore, we used Spearman's rank correlation coefficient to assess the relationships between other sleep features and IG values. These analyses were conducted using SciPy, and results were visualized with Matplotlib in Figure 5.8.

Figure 5.8 shows negative correlation between WASO, N3- Epochs and #WB, while all the other sleep features calculated have a positive correlation with IG values. This agrees with earlier works about N3 relationship with IG (Herzog et al., 2013; Tasali et al., 2008)

(decreased SWS increased IG levels), similarly increased WASO and WB which represent sleep suppression increase IG values (Leproult & Van Cauter, 2010). Other parameters have been shown to affect sleep in a complex manner that is not recorded in this simple test (for example, TST has a positive correlation with IG values up to a certain time and then has a negative correlation with IG values)

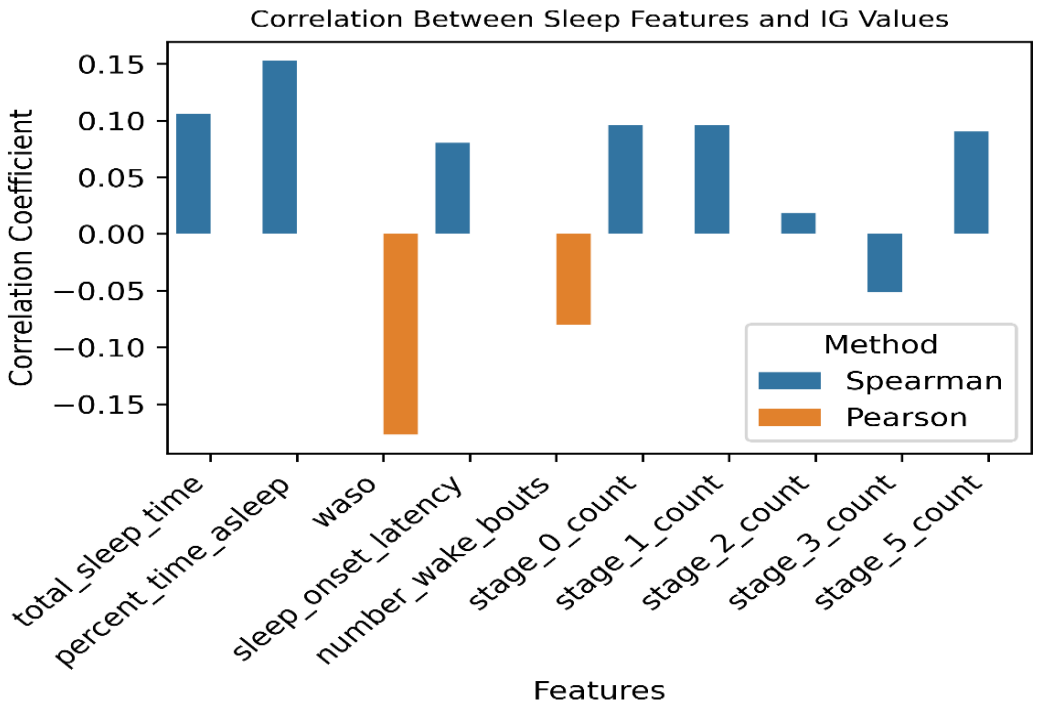


Figure 5.8: Correlation of sleep features with IG values.

To improve ML model performance, various statistical transformations—Log, Box-Cox, Yeo-Johnson, and Rank-Based Inverse Normal—are applied to the sleep features. The log transformation reduces skewness by compressing large values, making distributions more symmetric, but it requires all values to be positive. The Box-Cox transformation stabilizes variance and normalizes data through an optimal power transformation, but it is also restricted to positive values. Yeo-Johnson extends Box-Cox by accommodating zero and negative values, providing greater flexibility in normalizing data and stabilizing variance. The Rank-Based Inverse Normal transformation ranks the data and maps it to the quantiles of a normal distribution, reshaping the data to approximate normality.

We apply log, Box-Cox, Yeo-Johnson, and Rank-Based Inverse Normal transformations to normalize features and evaluate their correlation with IG values using Pearson correlation. After calculating correlations for both the original and transformed features, only those with significant correlation increase are selected for inclusion in the modified feature set. Based on the correlation with IG values the transformations marked with a tick sign in Figure 5.9 are selected

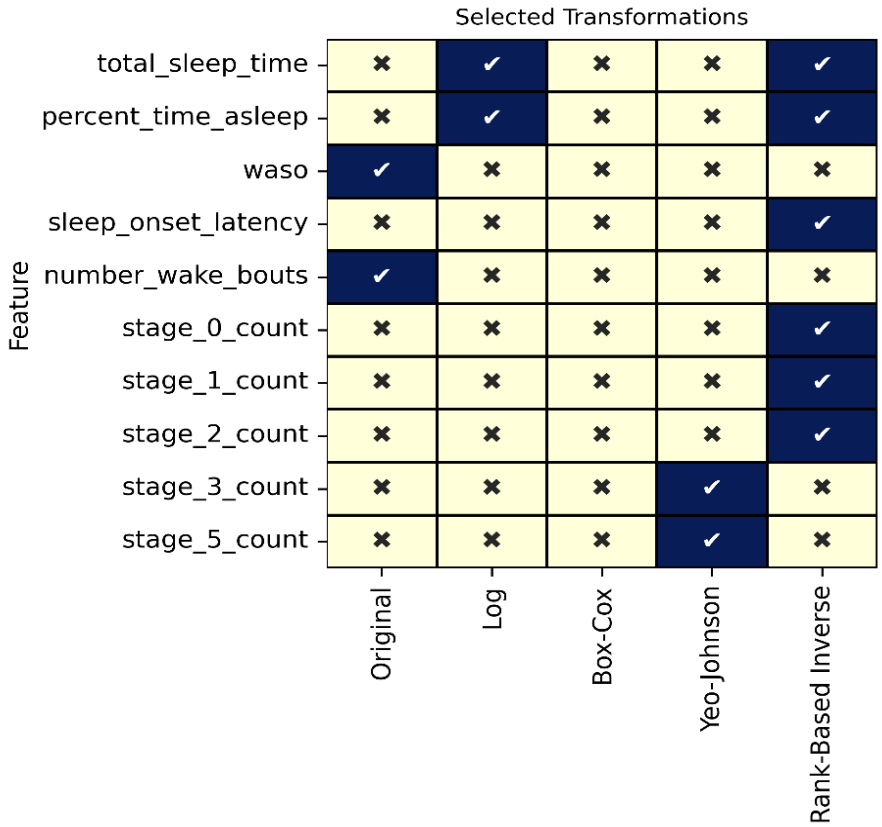


Figure 5.9: Selected transformations based on correlations with IG

The following additional features are defined: the percentage of time spent in each sleep stage (N1-Epoch-Percent, N2-Epoch-Percent, N3-Epoch-Percent, Wake-Epoch-Percent, REM-Epoch-Percent). Shapiro-Wilk test results indicate that none of these features are normally distributed. Spearman correlation analysis shows significant relationships for all of them as shown in Figure 5.10.

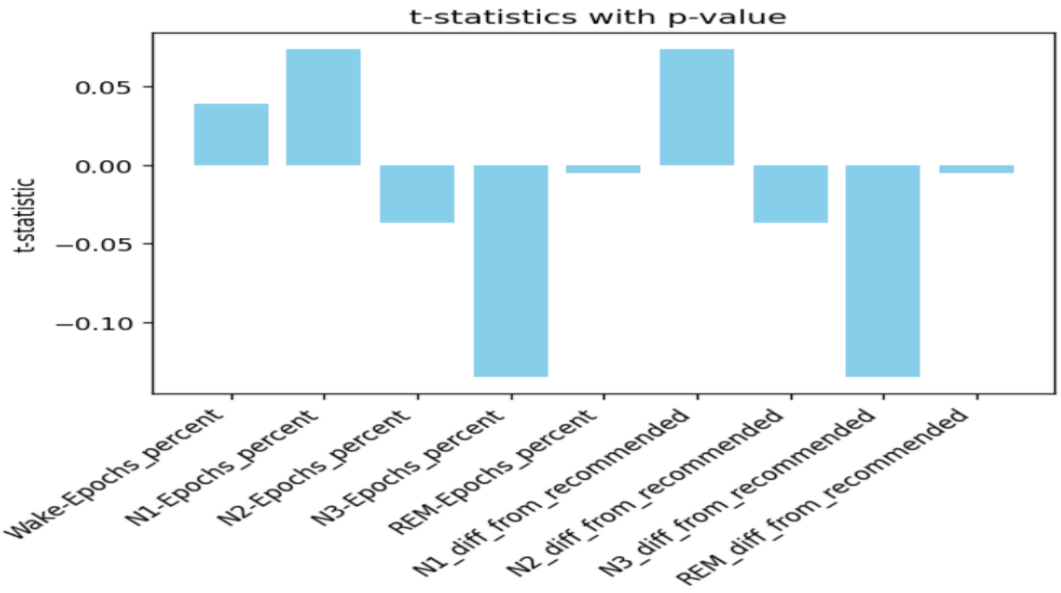


Figure 5.10: Correlations of new defined sleep features with IG values

Another set of sleep features is created by calculating the deviation of each sleep stage percentage from recommended values, based on guidelines from the American Academy of Sleep Medicine (AASM). According to AASM guidelines, the number of wake epochs should fall within 5-10%, so a midpoint of 7.5% is used to define feature called difference in wake epochs from recommended levels (Wake-Epochs_percent_diff_from_recommended). Similarly, N1 sleep stage epochs should be between 5-10%, so 7.5% is used to define a feature called difference in N1 epochs from recommended levels(N1-Epochs_percent_diff_from_recommended). For N2 sleep stages, the recommended range is 40-50%, and the average of 45% is used to define N2-Epochs_percent_diff_from_recommended. N3 sleep stages should occupy 20-25%, leading to an average of 22.5% for N3-Epochs_percent_diff_from_recommended. Finally, REM sleep stages should range from 20-25%, so 22.5% is used to define REM-Epochs_percent_diff_from_recommended.

The corresponding features are not normally distributed, hence their correlations with IG are measured using Spearman correlation and are all are statistically significant (shown in Figure 5.10). Additionally, the time stamps are separated into hours, minutes, and seconds, sine and cosine of these values are used as features. It is worth noting that correlations are a measure of linearity (or monotonicity in the case of Spearman)

5.6 Machine Learning models for IG prediction

To evaluate ML model performance for predicting IG values (regression), models are trained and validated using 10-fold cross-validation. The best models are further trained on 70% of the data and tested on the remaining 30%.

Table 5.7 represents different feature classes used in this work and what they mean

Table 5.7: Different Classes of Sleep Features used in this work

Feature Class	Explanation
Others	Defined in Chapter 3, and 4 (Ali et al., 2023; Bent, Cho, Henriquez, et al., 2021)
All Sleep Features	The sleep features include TST, PST, WASO, SOL and #WB. Each sleep stage, 0 (W) , 1 (N1), 2 (N2), 3 (N3), and 5 (REM), counts. Log transformations (i.e., stage_0_count_log, stage_2_countlog, total_sleep_time_log, wasolog,), Box-Cox and Yeo-Johnson transformations (i.e., stage_2_count_boxcox, stage_3_count_yeojohnson), and rank based inverse transforms (total_sleep_time_rank, waso_rank).
Transformed Sleep Features	The transformed sleep features various log, rank, and percentage transformations defined in Figure 5.9 as well as TST, PST, WASO, SOL and #WB. Each sleep stage, 0 (W), 1 (N1), 2 (N2), 3 (N3), and 5 (REM). Log transformations include `stage_0_count_log`, `total_sleep_time_log`, and `percent_time_asleep_log`, which normalize the distribution of sleep stages and metrics. Rank based inverse transform are applied to metrics like `total_sleep_time_rank`, `percent_time_asleep_rank`, and `sleep_onset_latency_rank. The `Yeo-Johnson` transformation, applied to `stage_3_count` and `stage_5_count`, also adjusts data skewness. Additionally, percentages for each sleep stage with further features indicating how they deviate from recommended values (e.g., stage_3_count_percent_diff_from_recommended).
Unmodified Sleep features	The unmodified sleep features include TST, PST, WASO, SOL. Counts for specific sleep stages are also included in unmodified sleep features.

5.6.1 Regression

The dataset features are split into other features (identified in Chapter 3), unmodified sleep features, and modified sleep features (Explained in Table 5.7). Models are trained on these combinations: (1) other features, (2) other + unmodified sleep features, (3) other + modified sleep features, and (4) other + all sleep features. Table 5.8 shows the results of 10-fold cross validation for the regression performance metrics. Figure 5.11 shows the MAE values for different models highlighting the performance of tree-based models (RF, DT and XGBOOST).

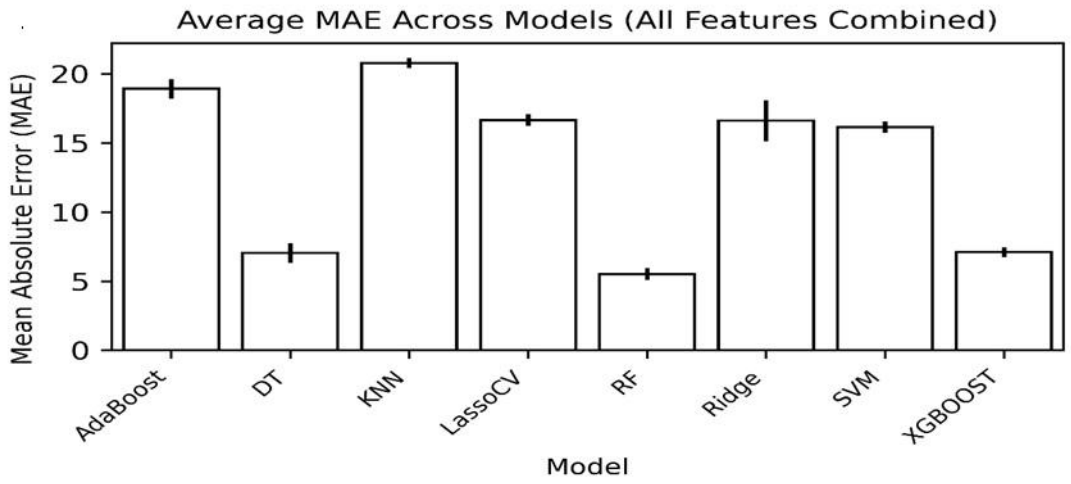


Figure 5.11: Comparison of MAE for different ML models for predicting IG values

Table 5.8: ***Comparison of models for predicting IG values***

Model	Feature Set	MAE	MAPE	R2	Adjusted R2	MSLE	Ev	RMSE
AdaBoost	other_features	19.396 ± 0.745	18.036 ± 0.782	-0.084 ± 0.08	-0.129 ± 0.083	23.483 ± 0.557	0.168 ± 0.025	23.483 ± 0.557
	other_features + all_sleep_features	18.795 ± 0.676	17.39 ± 0.8	-0.023 ± 0.076	-0.104 ± 0.082	22.812 ± 0.602	0.195 ± 0.027	22.812 ± 0.602
	other_features + transformed_sleep_features	18.68 ± 0.56	17.282 ± 0.662	-0.012 ± 0.073	-0.068 ± 0.077	22.683 ± 0.438	0.2 ± 0.028	22.683 ± 0.438
	other_features + unmodified_sleep_features	18.799 ± 0.741	17.397 ± 0.759	-0.021 ± 0.06	-0.068 ± 0.062	22.794 ± 0.65	0.195 ± 0.034	22.794 ± 0.65
DT	other_features	8.135 ± 0.344	6.984 ± 0.335	0.576 ± 0.05	0.559 ± 0.053	14.658 ± 0.7	0.577 ± 0.05	14.658 ± 0.7
	other_features + all_sleep_features	6.605 ± 0.343	5.636 ± 0.258	0.71 ± 0.047	0.687 ± 0.051	12.106 ± 0.902	0.71 ± 0.047	12.106 ± 0.902
	other_features + transformed_sleep_features	6.607 ± 0.226	5.63 ± 0.167	0.715 ± 0.038	0.699 ± 0.04	12.025 ± 0.688	0.715 ± 0.038	12.025 ± 0.688
	other_features + unmodified_sleep_features	6.829 ± 0.29	5.845 ± 0.243	0.684 ± 0.04	0.669 ± 0.042	12.672 ± 0.745	0.684 ± 0.04	12.672 ± 0.745
KNN	other_features	20.99 ± 0.347	17.991 ± 0.383	-0.609 ± 0.094	-0.675 ± 0.098	28.607 ± 0.401	-0.585 ± 0.096	28.607 ± 0.401
	other_features + all_sleep_features	20.609 ± 0.339	17.65 ± 0.368	-0.573 ± 0.098	-0.698 ± 0.105	28.291 ± 0.414	-0.547 ± 0.099	28.291 ± 0.414
	other_features + transformed_sleep_features	20.955 ± 0.338	17.961 ± 0.374	-0.606 ± 0.094	-0.696 ± 0.099	28.589 ± 0.398	-0.582 ± 0.096	28.589 ± 0.398
	other_features + unmodified_sleep_features	20.615 ± 0.345	17.655 ± 0.372	-0.574 ± 0.098	-0.647 ± 0.103	28.297 ± 0.422	-0.548 ± 0.099	28.297 ± 0.422
LassoCV	other_features	16.663 ± 0.442	14.458 ± 0.4	-0.001 ± 0.001	-0.042 ± 0.001	22.592 ± 0.696	-0.0 ± 0.0	22.592 ± 0.696
	other_features + all_sleep_features	16.663 ± 0.442	14.458 ± 0.4	-0.001 ± 0.001	-0.08 ± 0.001	22.592 ± 0.696	-0.0 ± 0.0	22.592 ± 0.696
	other_features + transformed_sleep_features	16.663 ± 0.442	14.458 ± 0.4	-0.001 ± 0.001	-0.057 ± 0.001	22.592 ± 0.696	-0.0 ± 0.0	22.592 ± 0.696
	other_features + unmodified_sleep_features	16.663 ± 0.442	14.458 ± 0.4	-0.001 ± 0.001	-0.048 ± 0.001	22.592 ± 0.696	-0.0 ± 0.0	22.592 ± 0.696
RF	other_features	6.243 ± 0.121	5.343 ± 0.157	0.815 ± 0.012	0.807 ± 0.012	9.712 ± 0.265	0.815 ± 0.012	9.712 ± 0.265

	other_features + all_sleep_features	5.267 ± 0.133	4.478 ± 0.155	0.866 ± 0.011	0.855 ± 0.012	8.264 ± 0.244	0.866 ± 0.011	8.264 ± 0.244
	other_features + transformed_sleep_features	5.271 ± 0.125	4.48 ± 0.132	0.865 ± 0.013	0.857 ± 0.014	8.282 ± 0.289	0.865 ± 0.013	8.282 ± 0.289
	other_features + unmodified_sleep_features	5.33 ± 0.16	4.526 ± 0.166	0.864 ± 0.012	0.857 ± 0.013	8.324 ± 0.282	0.864 ± 0.012	8.324 ± 0.282
Ridge	other_features	17.158 ± 1.946	14.86 ± 1.698	-0.084 ± 0.247	-0.128 ± 0.257	23.388 ± 2.604	-0.083 ± 0.247	23.388 ± 2.604
	other_features + all_sleep_features	16.134 ± 0.811	13.963 ± 0.812	0.036 ± 0.08	-0.04 ± 0.087	22.146 ± 0.951	0.037 ± 0.081	22.146 ± 0.951
	other_features + transformed_sleep_features	16.743 ± 1.396	14.499 ± 1.363	-0.031 ± 0.136	-0.088 ± 0.144	22.887 ± 1.801	-0.03 ± 0.137	22.887 ± 1.801
	other_features + unmodified_sleep_features	16.436 ± 1.58	14.241 ± 1.547	-0.004 ± 0.195	-0.05 ± 0.204	22.517 ± 1.787	-0.002 ± 0.195	22.517 ± 1.787
SVM	other_features	16.149 ± 0.421	13.66 ± 0.355	0.018 ± 0.009	-0.023 ± 0.01	22.38 ± 0.67	0.034 ± 0.007	22.38 ± 0.67
	other_features + all_sleep_features	16.149 ± 0.421	13.66 ± 0.355	0.018 ± 0.009	-0.06 ± 0.01	22.38 ± 0.67	0.034 ± 0.007	22.38 ± 0.67
	other_features + transformed_sleep_features	16.149 ± 0.421	13.66 ± 0.355	0.018 ± 0.009	-0.037 ± 0.01	22.38 ± 0.67	0.034 ± 0.007	22.38 ± 0.67
	other_features + unmodified_sleep_features	16.149 ± 0.421	13.66 ± 0.355	0.018 ± 0.009	-0.028 ± 0.01	22.38 ± 0.67	0.034 ± 0.007	22.38 ± 0.67
XGBOOST	other_features	7.657 ± 0.147	6.575 ± 0.139	0.769 ± 0.015	0.759 ± 0.016	10.842 ± 0.281	0.769 ± 0.016	10.842 ± 0.281
	other_features + all_sleep_features	6.898 ± 0.132	5.925 ± 0.14	0.812 ± 0.009	0.797 ± 0.01	9.787 ± 0.241	0.812 ± 0.009	9.787 ± 0.241
	other_features + transformed_sleep_features	6.9 ± 0.147	5.922 ± 0.138	0.81 ± 0.014	0.799 ± 0.015	9.84 ± 0.256	0.81 ± 0.014	9.84 ± 0.256
	other_features + unmodified_sleep_features	6.98 ± 0.146	5.987 ± 0.176	0.809 ± 0.015	0.8 ± 0.016	9.865 ± 0.19	0.809 ± 0.015	9.865 ± 0.19

For better performance of the models the RMSE, MAE, MSLE and MAPE should be low, thus implying that RF, DT and XGBOOST are the best performing models. In table 5.8 it can be seen that the MAE of DT is reduced from 8.135 ± 0.344 mg/dL to 6.829 ± 0.29 mg/dL highlighting the performance increase by adding sleep features.

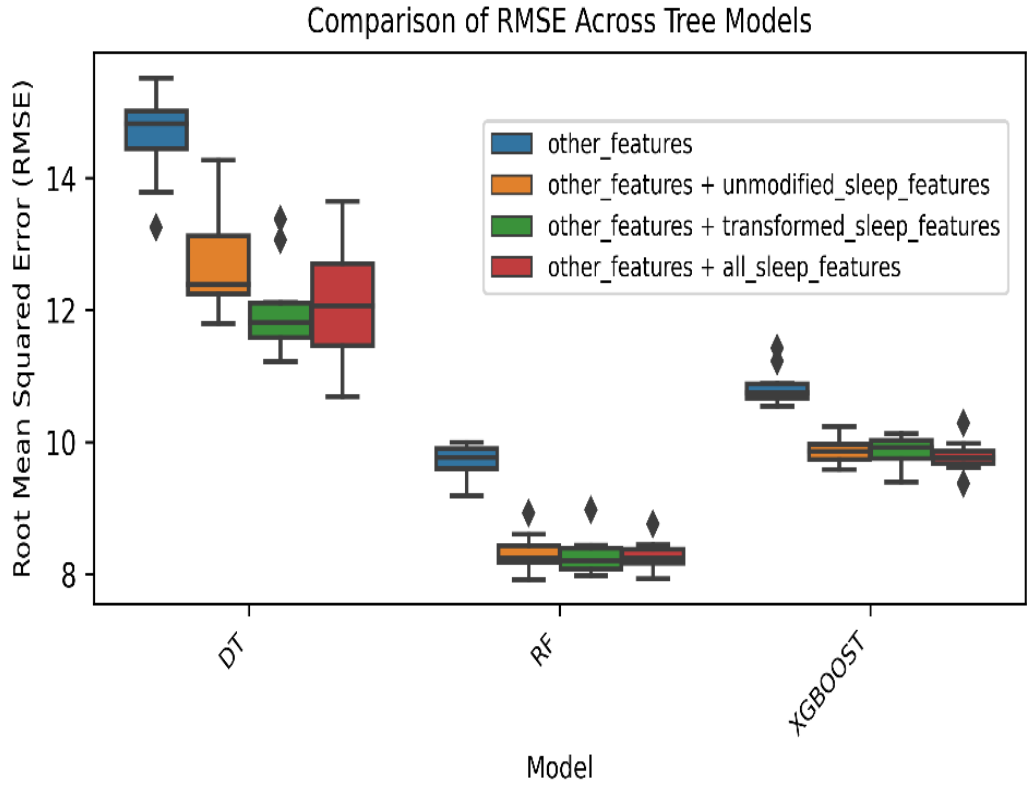


Figure 5.12: Comparison of RMSE for all feature sets with tree models

This comparison highlights the reduction in RMSE with the inclusion of sleep features, and further improvements with the modification of those features. The performance gains from adding sleep features are consistent across all models, with DT and RF showing the most significant improvements after sleep feature modification. Wilcoxon paired t-tests were conducted separately for each model to assess the statistical significance of these improvements. The Wilcoxon t-tests for DT and RF specifically confirm the significant performance boost from adding and modifying sleep features, as shown in Figure 5.13.

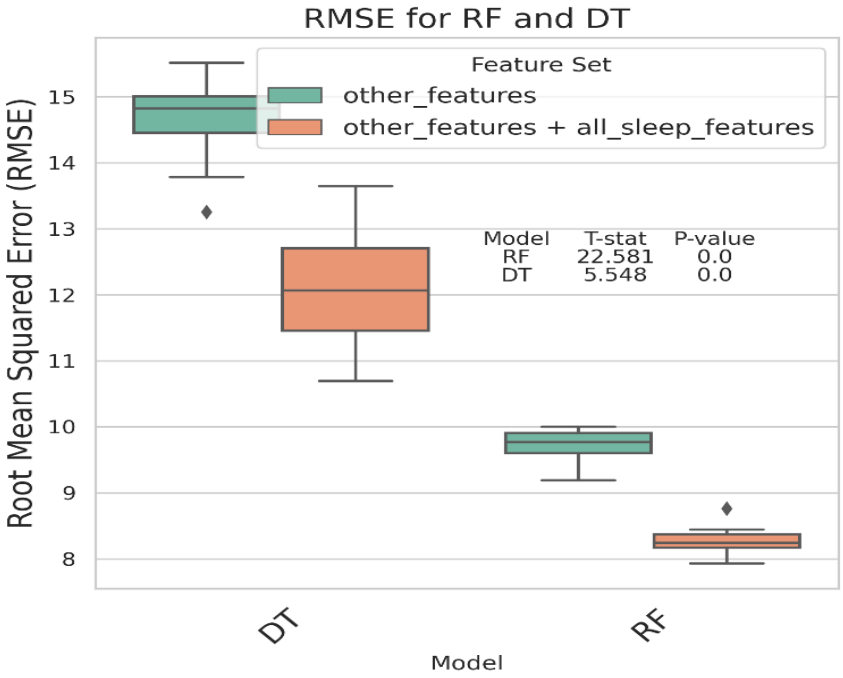


FIGURE 5.13: Paired t-tests for RF and DT RMSE values with and without the addition of sleep features

After comparing the expected values of different error metrics, we compare all models across all feature combinations using Friedman's test and Nemenyi Post Hoc analysis (Figure 5.14).

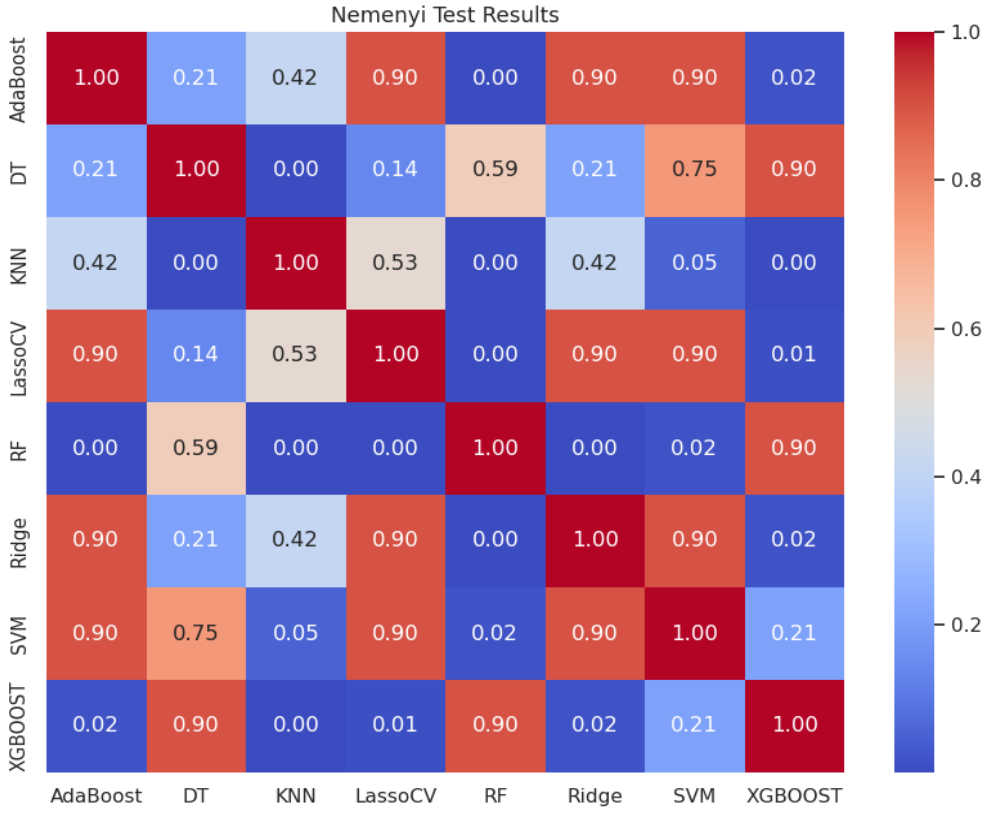


Figure 5.14: Nemenyi Post hoc results for RMSE values of all the models using all features

From the heatmap, we can observe that models like RF and XGBOOST show significant differences in performance compared to other models, particularly AdaBoost and KNN. Conversely, comparisons like LassoCV versus Ridge and DT versus SVM do not show significant differences, indicating similar performance classes between these models. After this analysis, it can be concluded that RF is significantly better than other models.

To find the feature importances for different data sources, we find impurity-based feature importances across all folds and average them resulting in the following relative importances (Figure 5.15).

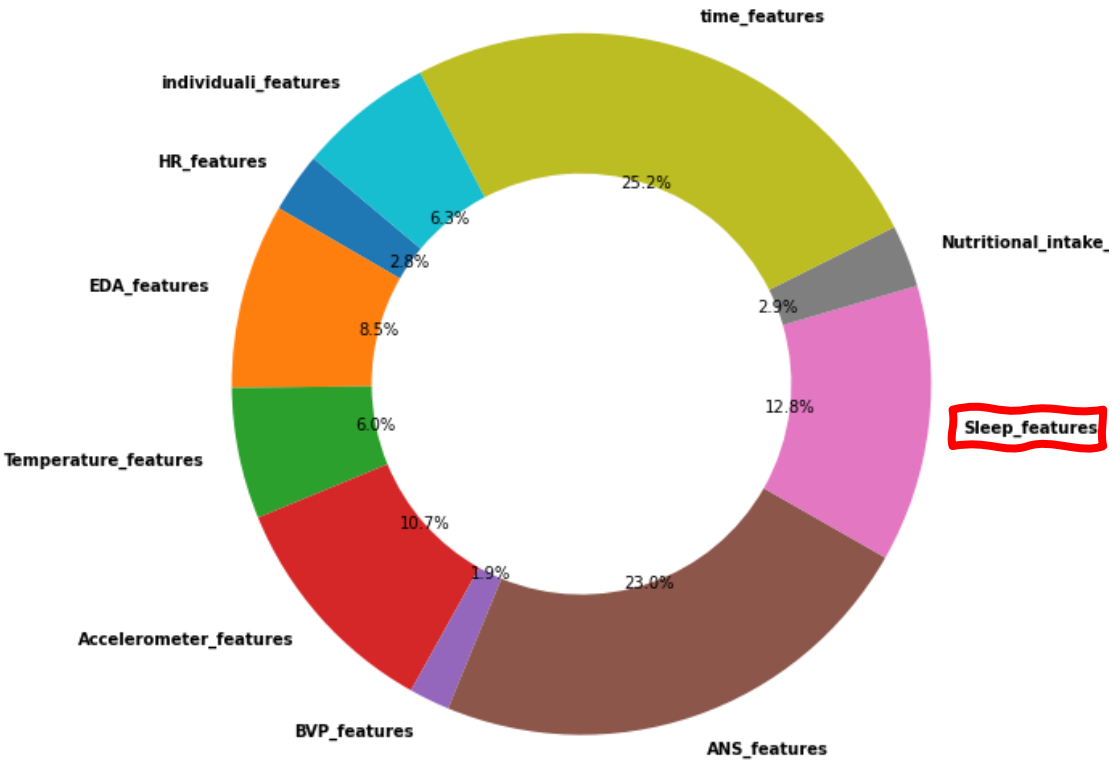


Figure 5.15: Relative feature importance for different feature types

5.6.2 Classification

Before classifying IG values into high, low, and normal (as shown in Figure 5.2), the data was resampled to ensure 2,500 rows per class. The models were trained and evaluated using 10-fold cross-validation. A summary of the models' performance is provided in confusion matrices (Figure 5.16 and Figure 51.17).

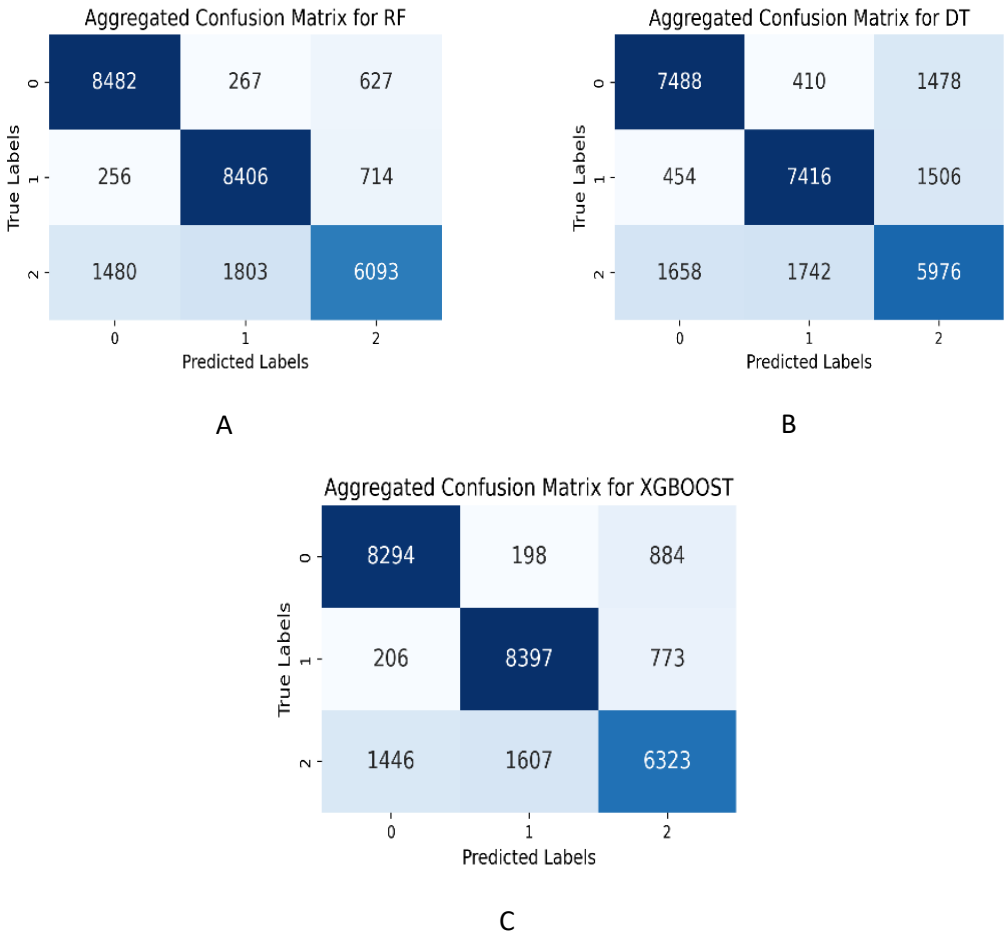


Figure 5.16*:* Confusion matrix for tree models (0= Low Glucose, 1=Normal Glucose and 2= High Glucose) A-Confusion matrix of RF model B- Confusion matrix of DT model C-Confusion matrix of XGBOOST model

From the increased prediction accuracy of tree models (Table 5.9), it can be observed adding sleep features have improved the prediction performance of the models. For the comparison amongst models the performance of each fold is aggregated for all the feature sets. The aggregated confusion matrices are given in Figure 5.16 for tree models. Note that the sum of rows exceeds 2500 because this is aggregated for all the feature sets (others, others+ unmodified sleep features, other + modified sleep features and others +all sleep features). The confusion matrices of the tree models shown in Figure 5.16, show the superior performance of tree models in the classification problem as well, but the slightly less support of the high glucose shows that the feature representation can be improved to enhanced by increasing the distance amongst classes. Confusion matrix of non-tree classifiers in Figure 5.17 show the poor performance of non-tree classifiers for IG level classification. This performance can be tentatively increased by selecting the right transformations on all the features that mitigate the effects of outliers and scales.

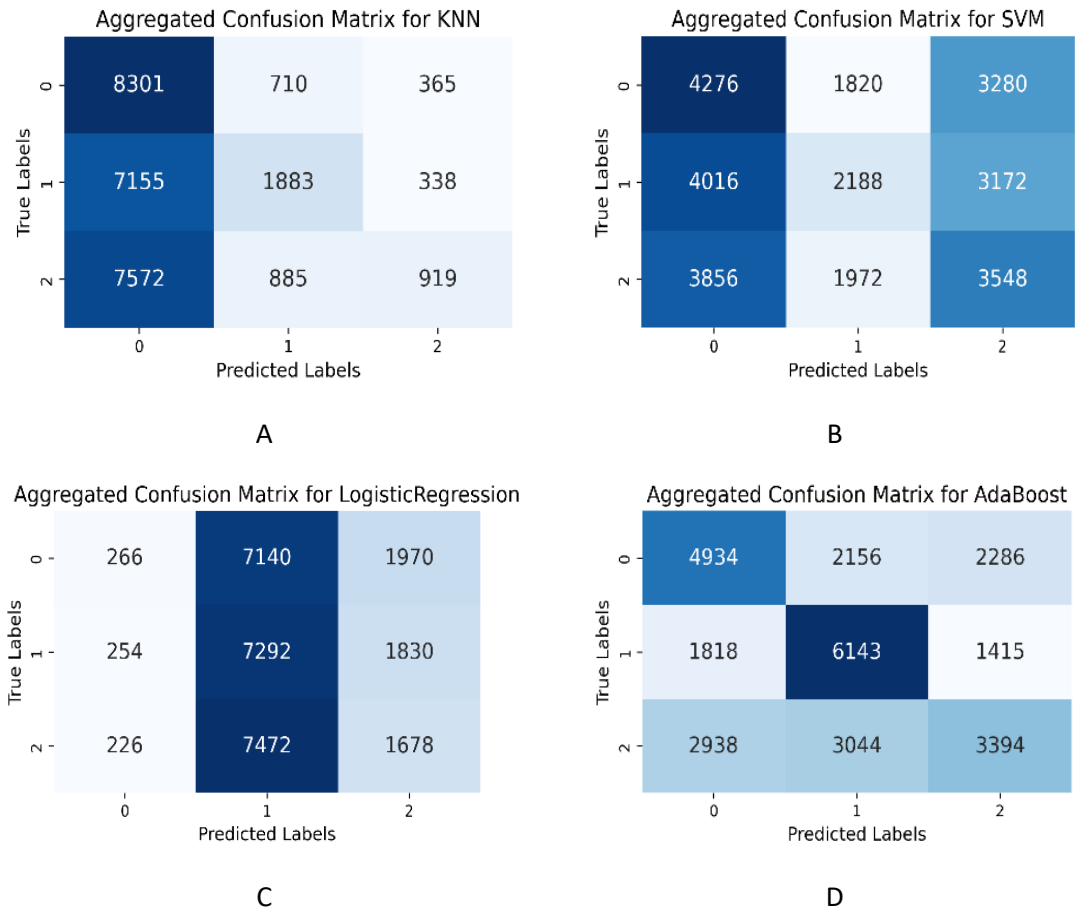


Figure 5.17: Confusion matrix for non-tree classifiers. A-Confusion matrix for KNN, B-Confusion matrix for SVM, C- Confusion matrix for ADABOOST, D- Confusion matrix for Logistic regression.

Confusion matrix of non-tree classifiers in Figure 5.17 show the poor performance of non-tree classifiers on this task except for Adaboost which is also a tree classifier.

From the accuracy comparison of different models in Figure 5.18, it can be concluded that tree models (RF, DT and XGBOOST) outperform other models in classification tasks as well.

Table 5.9 compares the performance of different ML models in classification task.

Table 5.9: Comparison of ML models for classification of IG values into personalized high, normal and low glucose and the effect of sleep on performance of ML models.

Model	Feature Set	Accuracy (mean ± std)	F1 Score (mean ± std)	Precision (mean ± std)	Recall (mean ± std)
AdaBoost	other_features	0.519 ± 0.017	0.512 ± 0.018	0.518 ± 0.019	0.519 ± 0.017
	other_features + all_sleep_features	0.515 ± 0.026	0.509 ± 0.027	0.513 ± 0.027	0.515 ± 0.026
	other_features + transformed_sleep_features	0.517 ± 0.025	0.51 ± 0.027	0.515 ± 0.028	0.517 ± 0.025
	other_features + unmodified_sleep_features	0.507 ± 0.017	0.499 ± 0.018	0.503 ± 0.017	0.507 ± 0.017
DT	other_features	0.715 ± 0.02	0.713 ± 0.021	0.713 ± 0.021	0.715 ± 0.02
	other_features + all_sleep_features	0.751 ± 0.012	0.75 ± 0.013	0.75 ± 0.013	0.751 ± 0.012
	other_features + transformed_sleep_features	0.751 ± 0.015	0.75 ± 0.016	0.75 ± 0.016	0.751 ± 0.015
	other_features + unmodified_sleep_features	0.752 ± 0.01	0.752 ± 0.01	0.752 ± 0.011	0.752 ± 0.01
KNN	other_features	0.389 ± 0.025	0.317 ± 0.03	0.478 ± 0.034	0.389 ± 0.025
	other_features + all_sleep_features	0.399 ± 0.024	0.331 ± 0.03	0.503 ± 0.039	0.399 ± 0.024
	other_features + transformed_sleep_features	0.39 ± 0.025	0.318 ± 0.03	0.482 ± 0.034	0.39 ± 0.025
	other_features + unmodified_sleep_features	0.399 ± 0.024	0.331 ± 0.029	0.504 ± 0.037	0.399 ± 0.024
LR	other_features	0.328 ± 0.01	0.238 ± 0.011	0.215 ± 0.014	0.328 ± 0.01
	other_features + all_sleep_features	0.329 ± 0.009	0.24 ± 0.013	0.229 ± 0.05	0.329 ± 0.009
	other_features + transformed_sleep_features	0.328 ± 0.01	0.238 ± 0.011	0.215 ± 0.014	0.328 ± 0.01
	other_features + unmodified_sleep_features	0.329 ± 0.009	0.24 ± 0.013	0.229 ± 0.05	0.329 ± 0.009
RF	other_features	0.799 ± 0.018	0.794 ± 0.019	0.8 ± 0.019	0.799 ± 0.018
	other_features + all_sleep_features	0.827 ± 0.011	0.823 ± 0.012	0.827 ± 0.01	0.827 ± 0.011
	other_features + transformed_sleep_features	0.825 ± 0.016	0.82 ± 0.017	0.825 ± 0.016	0.825 ± 0.016
	other_features + unmodified_sleep_features	0.818 ± 0.015	0.814 ± 0.016	0.819 ± 0.015	0.818 ± 0.015
SVM	other_features	0.356 ± 0.014	0.349 ± 0.013	0.359 ± 0.015	0.356 ± 0.014
	other_features + all_sleep_features	0.356 ± 0.014	0.349 ± 0.013	0.359 ± 0.015	0.356 ± 0.014
	other_features + transformed_sleep_features	0.356 ± 0.014	0.349 ± 0.013	0.359 ± 0.015	0.356 ± 0.014
	other_features + unmodified_sleep_features	0.356 ± 0.014	0.349 ± 0.013	0.359 ± 0.015	0.356 ± 0.014
XGBOOST	other_features	0.809 ± 0.014	0.806 ± 0.014	0.808 ± 0.014	0.809 ± 0.014
	other_features + all_sleep_features	0.822 ± 0.013	0.819 ± 0.013	0.821 ± 0.013	0.822 ± 0.013
	other_features + transformed_sleep_features	0.823 ± 0.015	0.82 ± 0.015	0.822 ± 0.015	0.823 ± 0.015
	other_features + unmodified_sleep_features	0.818 ± 0.011	0.815 ± 0.011	0.817 ± 0.011	0.818 ± 0.011

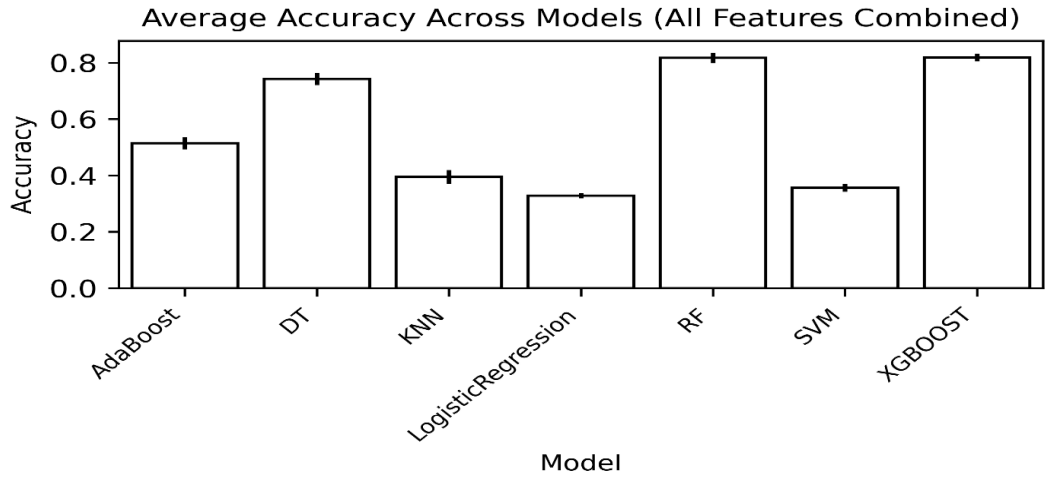


Figure 5.18: Accuracy comparison for classification models of IG classes

The accuracy improvement of the tree models by additional sleep features is in the Figure 5.19. This shows the increase in performance of classification models with additional sleep features, but it also shows that the statistical transformations have limited or no effect on the performance of the models.

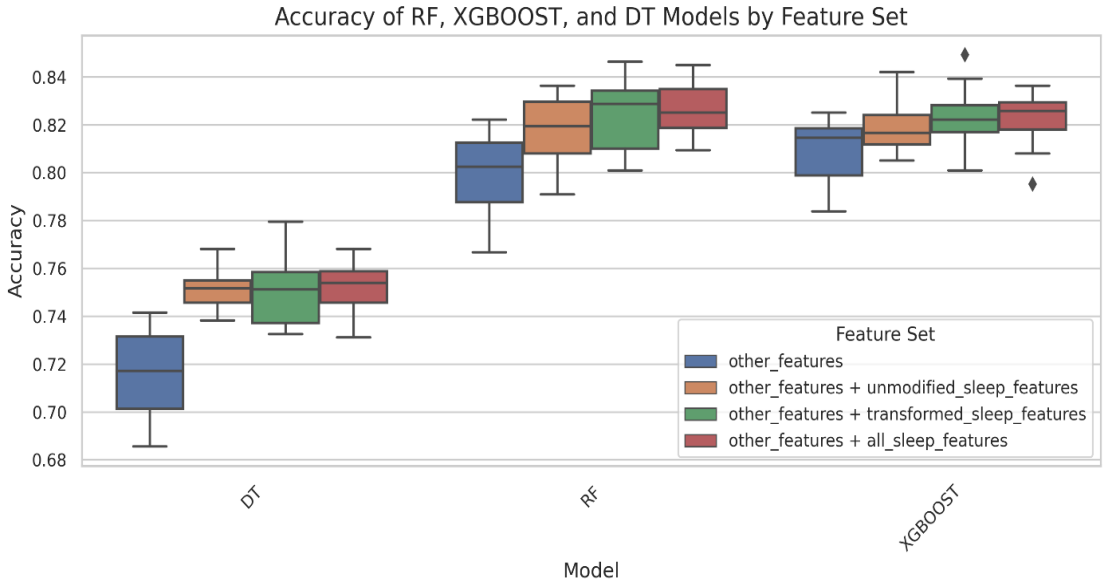


Figure 5.19: Accuracy comparison for tree-based classification models of IG classes

The paired t-tests for improved performance of the RF model are expanded in Figure 5.20 showing each of those performance increases are statistically significant. After ensuring the performance improvements across the models using the Friedman test, Nemenyi post hoc analysis is carried out and the results are plotted as a heatmap (Figure 5.21).

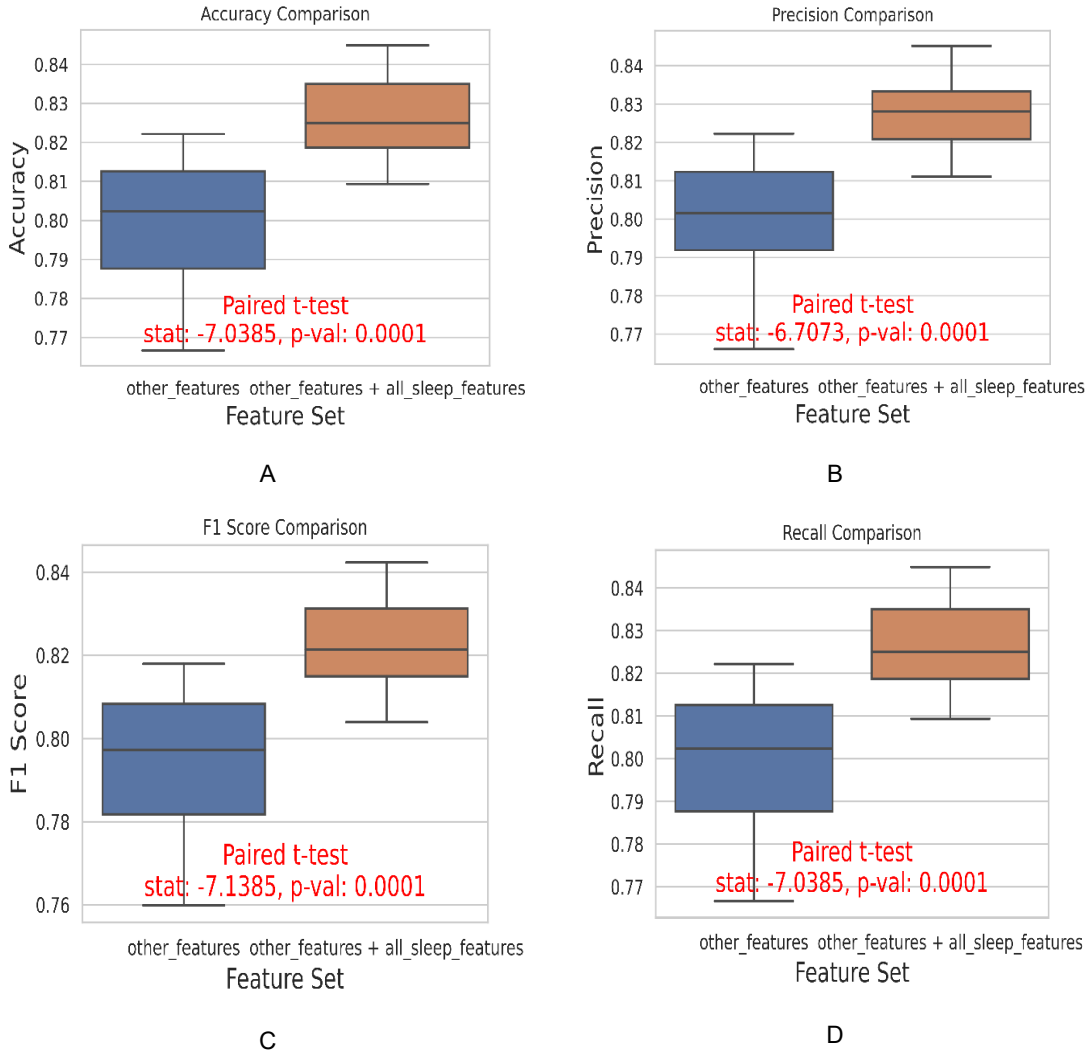


Figure 5.20: Performance increase by adding sleep features in an RF model. A- Increase in accuracy by adding modified sleep features, B- Increase in precision by adding modified sleep features. C- Increase in F1-score by adding modified sleep features. D- Increase in Recall by adding modified sleep features.

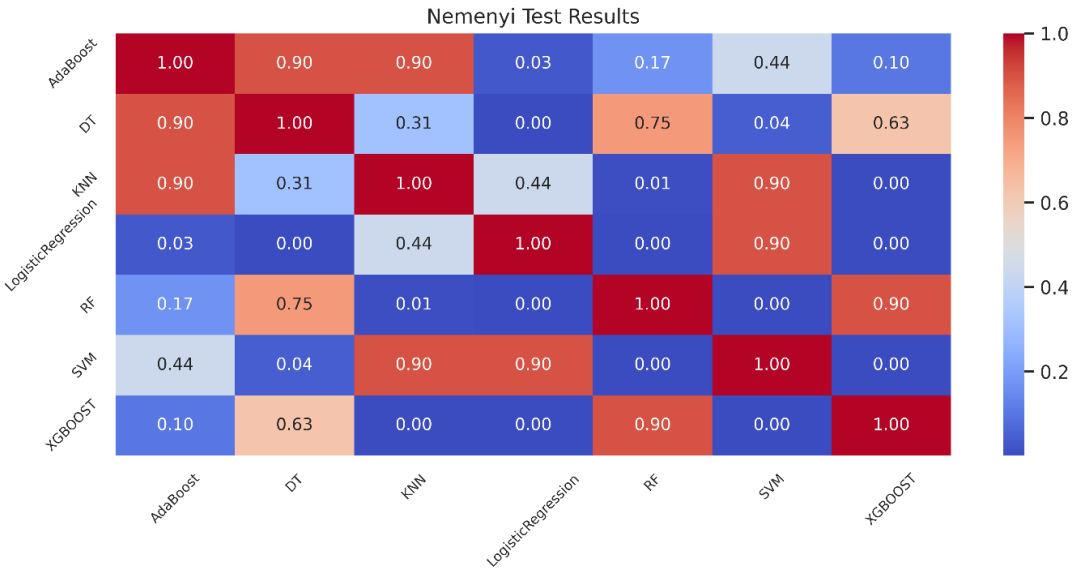


Figure 5.21: NEMENYI post hoc analysis to compare different models

This shows performance of the model XGBoost significantly improved over KNN, LR, and SVM). SVM is statistically different from DT, RF and XGBoost. RF is significantly different from KNN, LR and SVM. DT is significantly different from LR and SVM. Whereas Adaboost is significantly different from LR.

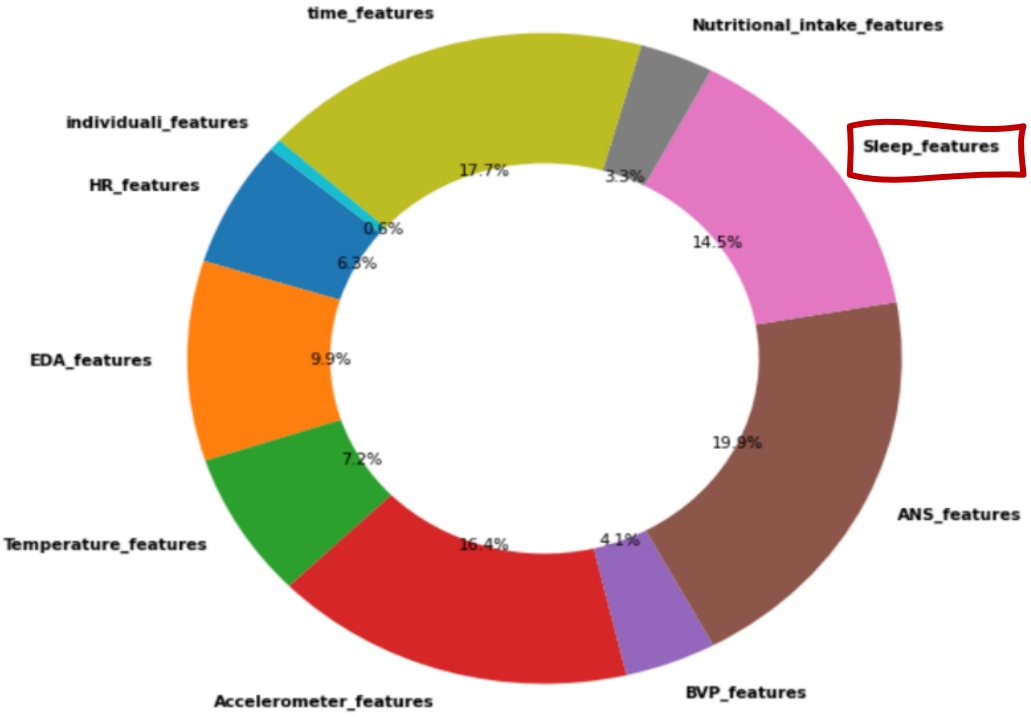
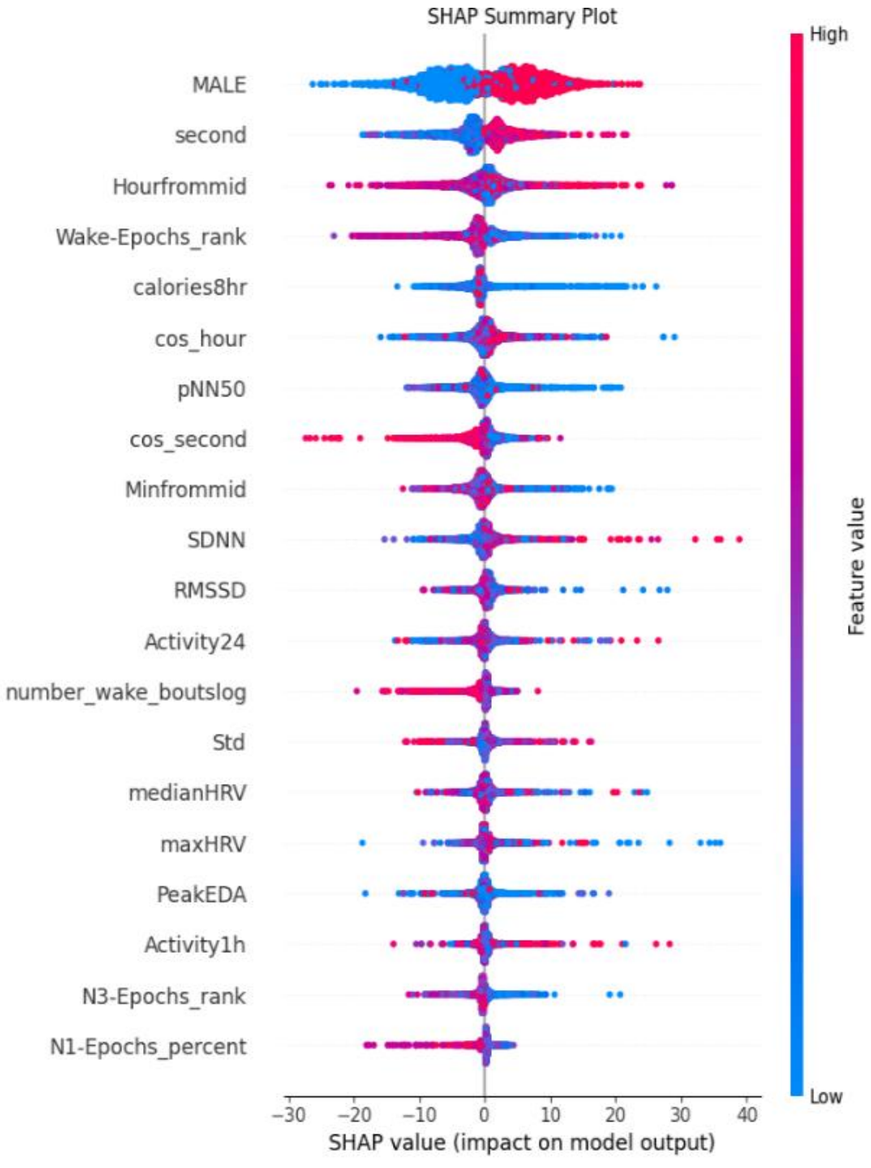


Figure 5.22: Relative feature importance in RF model for IG level classification

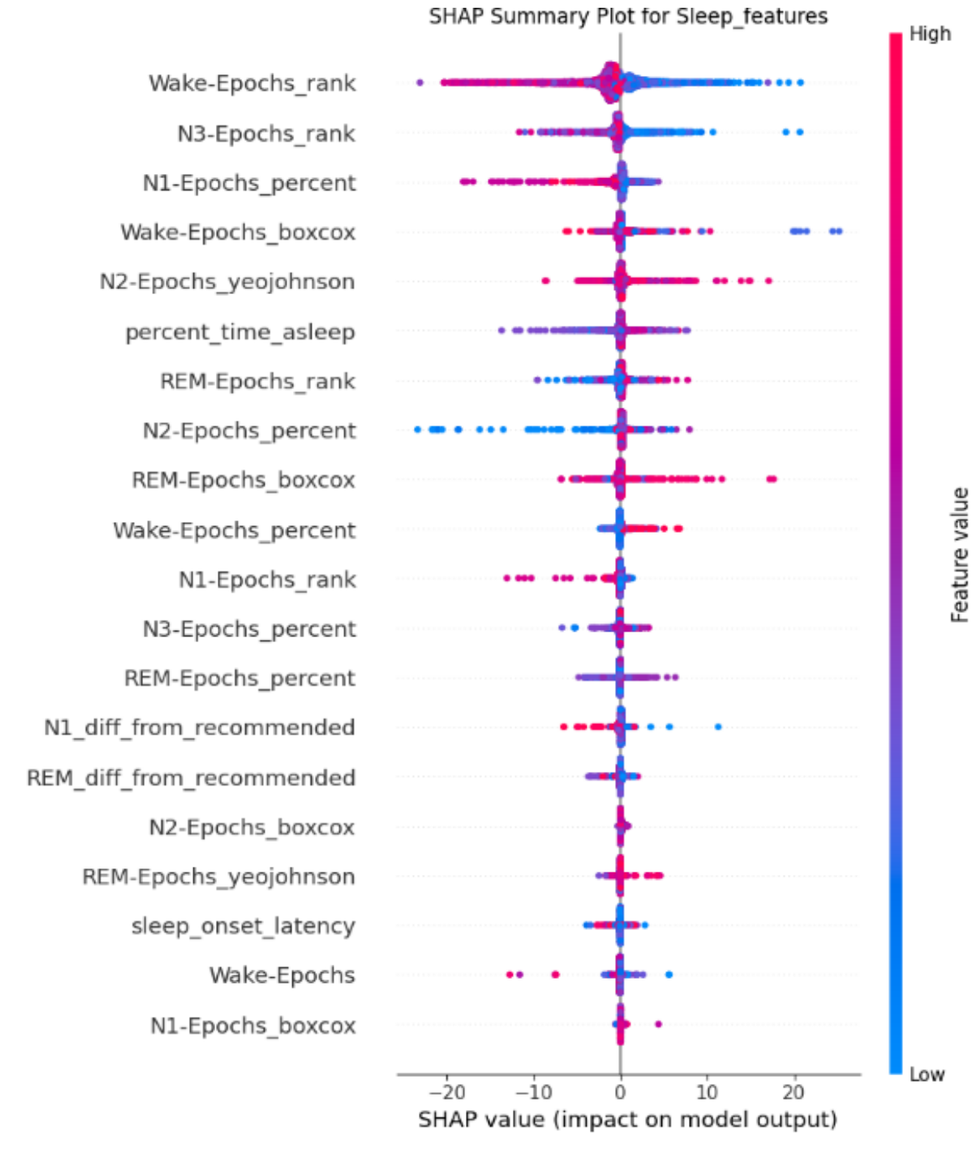
To find the feature importances for different data sources, we find impurity-based feature importances across all folds and average them resulting in the Figure 5.22. This shows that sleep features are 14.5 % important in comparison to the other features for an RF model in classification of glucose levels.

5.6.3 Model Explanations

RF models for IG prediction (regression and classification) are explained for all the feature sets using SHAP values. For this the models were trained using 70% data and tested using 30 % data. For the classification problem the number of classes in each category (high, normal, and low) were set to 2500. The model explanations are also plotted for different feature categories to see how the feature distributions effect the label prediction. The top features used for a RF regression model of IG values from all the input features (others+all sleep features) are shown in the SHAP plots below (Figure 5.23)



(a)



(b)

Figure 5.23: SHAP values for features in Prediction of Glucose levels for RF model trained on 70% data and tested on 30% data. (a) SHAP values of the top 20 features. (B) SHAP value for sleep features

SHAP values (Figure 5.23a) of the DT regression model shows the importance of circadian rhythm (second, hour from midnight, cos_hour , etc) in glucose prediction consistent earlier findings (Abbas et al., 2018; Bent, Cho, Henriquez, et al., 2021). Alongside these ANS based features such as meanHRV, SDNN and peakEDA values are also very significant predictors of IG values. The gender feature (male) is the most impactful feature, with high SHAP values spread across the range, suggesting that being male has a significant influence on the predicted IG values. Second feature represents the time component (seconds), and its importance suggests that the exact time of the day in seconds has a substantial effect on glucose classes. The oscillation in SHAP values across the feature's range shows a cyclical influence, possibly tied to circadian rhythms or mealtimes. Hourfrommid and Minfrommid representing time in hours and minutes from midnight, are also crucial. Their significant impact reflects the strong influence of the time of day on glucose regulation, which aligns with known circadian effects on metabolism. Wake-Epochs_rank and number_wake_boutslog features associated with sleep disturbances (e.g., wake bouts), show that frequent awakenings or disruptions in sleep have a notable effect on IG classes. This aligns with evidence that poor sleep quality can lead to impaired glucose metabolism. N3-Epochs_rank and N1-Epochs_percent features represent sleep stages, with N3 (deep sleep) and N1 (light sleep) contributing to glucose predictions. The impact suggests that different sleep stages, particularly deep sleep, may play a role in glucose regulation. SDNN, RMSSD, pNN50, medianHRV, maxHRV metrics are important for IG prediction, indicating that autonomic nervous system activity, which HRV reflects, is closely tied to glucose regulation. HRV features generally show that variations in autonomic activity have a meaningful influence on glucose classes, likely due to their relationship with stress and metabolic processes. Activity24 and Activity1h features, representing physical activity over different periods, are shown to impact IG values. Physical activity is known to enhance glucose uptake by muscles, which can lead to lower glucose classes, thus explaining the observed SHAP value patterns. PeakEDA, a measure of physiological stress, is also an impactful feature. High EDA values may be associated with stress responses that elevate glucose classes through hormonal pathways like cortisol release. cos_hour and cos_second are cosine transformations of the hour and second of the day, used to capture cyclical patterns (e.g., circadian rhythms). Their importance further supports the idea that time-related cycles strongly influence glucose classes. calories8hr represents caloric intake over the past 8 hours. Its impact on IG predictions suggests that recent food consumption has a direct effect on glucose classes, though it is not as dominant as some other physiological or time-based features.

Analysing only the effect of sleep features (Figure 5.24(b)) in RF model, Wake-Epochs_rank and Wake-Epochs_boxcox are the most impactful, with high SHAP values. The "Wake-Epochs" features, which measure the number of wake episodes during sleep, generally have positive SHAP values when high. This suggests that frequent or prolonged awakenings during the night are associated with increased IG values. Poor sleep quality, characterized by frequent wake episodes, is known to impair glucose regulation (Knutson et al., 2007). N3 sleep, or deep sleep, is crucial for restorative functions and glucose metabolism. Higher SHAP values for these features, particularly when N3 sleep is reduced (lower percent or rank), indicate a negative impact on IG classes (Tasali et al., 2008). This aligns with the understanding that insufficient deep sleep can lead to poorer glucose control, resulting in higher IG classes. N1 sleep, which is the lightest sleep stage, has a more complex relationship with IG values. The SHAP values for N1 sleep suggest that an increase in light sleep (higher N1 percent) may be associated with higher IG levels. This might reflect poorer sleep quality overall, as more time spent in light sleep can be indicative of sleep fragmentation or difficulty progressing into deeper sleep stages. N2 sleep (N2-Epochs_yeojohnson and N2-Epochs_percent), a stage of intermediate sleep, shows a mixed influence on IG predictions. Moderate SHAP values suggest that both too little and too much N2 sleep could impact glucose classes. Since N2 is a stable sleep stage, deviations from normal amounts may indicate disruptions that affect metabolic processes. Percent_time_asleep represents the overall sleep efficiency or the percentage of time spent asleep relative to the total time in bed. Higher SHAP values for reduced sleep efficiency suggest that less time spent asleep is associated with increased IG classes, highlighting the importance of getting sufficient, uninterrupted sleep. REM sleep (REM-Epochs_rank and REM-Epochs_percent) is critical for cognitive functions and stress regulation. The SHAP values show that a lower percentage or rank of REM sleep tends to increase IG levels. This aligns with research indicating that inadequate REM sleep can disrupt hormonal balance and glucose regulation. N1_diff_from_recommended and REM_diff_from_recommended measure deviations from the recommended amounts of N1 and REM sleep. Positive SHAP values for these deviations suggest that straying from optimal sleep patterns, particularly in light sleep and REM stages, can lead to elevated IG levels. Sleep Onset Latency feature represents the time it takes to fall asleep. The SHAP values indicate that longer sleep onset latency (difficulty falling asleep) is associated with higher IG levels. Difficulty falling asleep may reflect underlying stress or other factors that negatively impact glucose metabolism.

Similarly, SHAP summary plot for a RF model trained on 70% balanced data for IG level classification is given in Figure 5.24. The feature importances for classification for each data sources are plotted. The importance of sleep features is given in the picture below.

The SHAP visualizations are also plotted to gain insights into how different features impact the classification of IG classes into high, low, and normal classes. cos_hour representing the cosine transformation of the hour component, has the highest impact across all classes, especially high glucose (class 1) and normal glucose (Class 2). Hourfrommid and Minfrommid, representing time from midnight in hours and minutes, also show significant importance across classes. pNN50 is a heart rate variability (HRV) metric and quartiles (Q1G, Q3G) are particularly impactful, likely reflecting physiological responses that correlate with glucose classes.

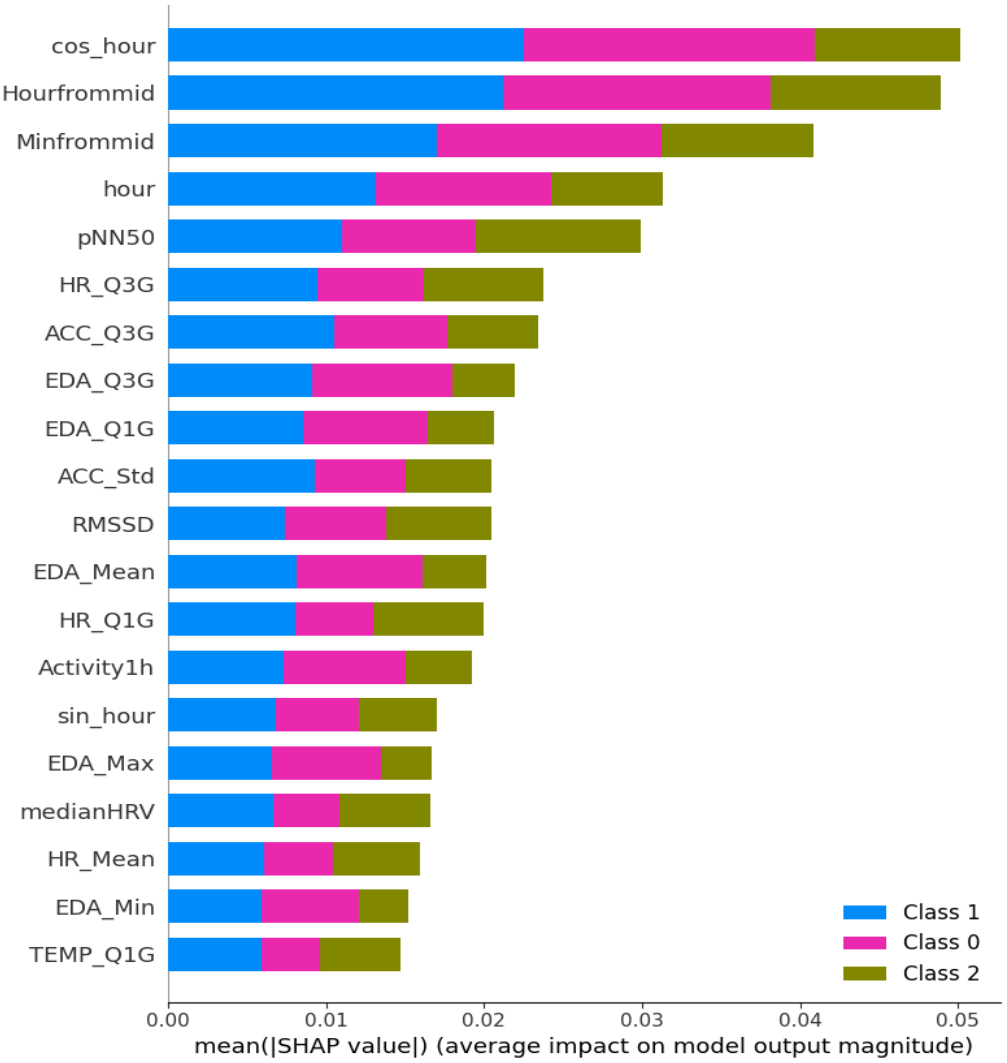
EDA and ACC features represent electrodermal activity and accelerometer data, respectively, and they also contribute to the classification, indicating that physical activity and stress may play roles in glucose fluctuations.

The SHAP values for various sleep-related features are also plotted. These values reveal that N3-Epochs (yeojohnson transformation), has a significant impact, particularly with higher values pushing the prediction towards high class. Sleep Onset Latency asleep has a moderate impact, with longer latencies slightly tilting the predictions. REM-Epochs Various transformations and rankings of REM sleep epochs contribute to the model, reflecting how REM sleep patterns may influence glucose regulation.

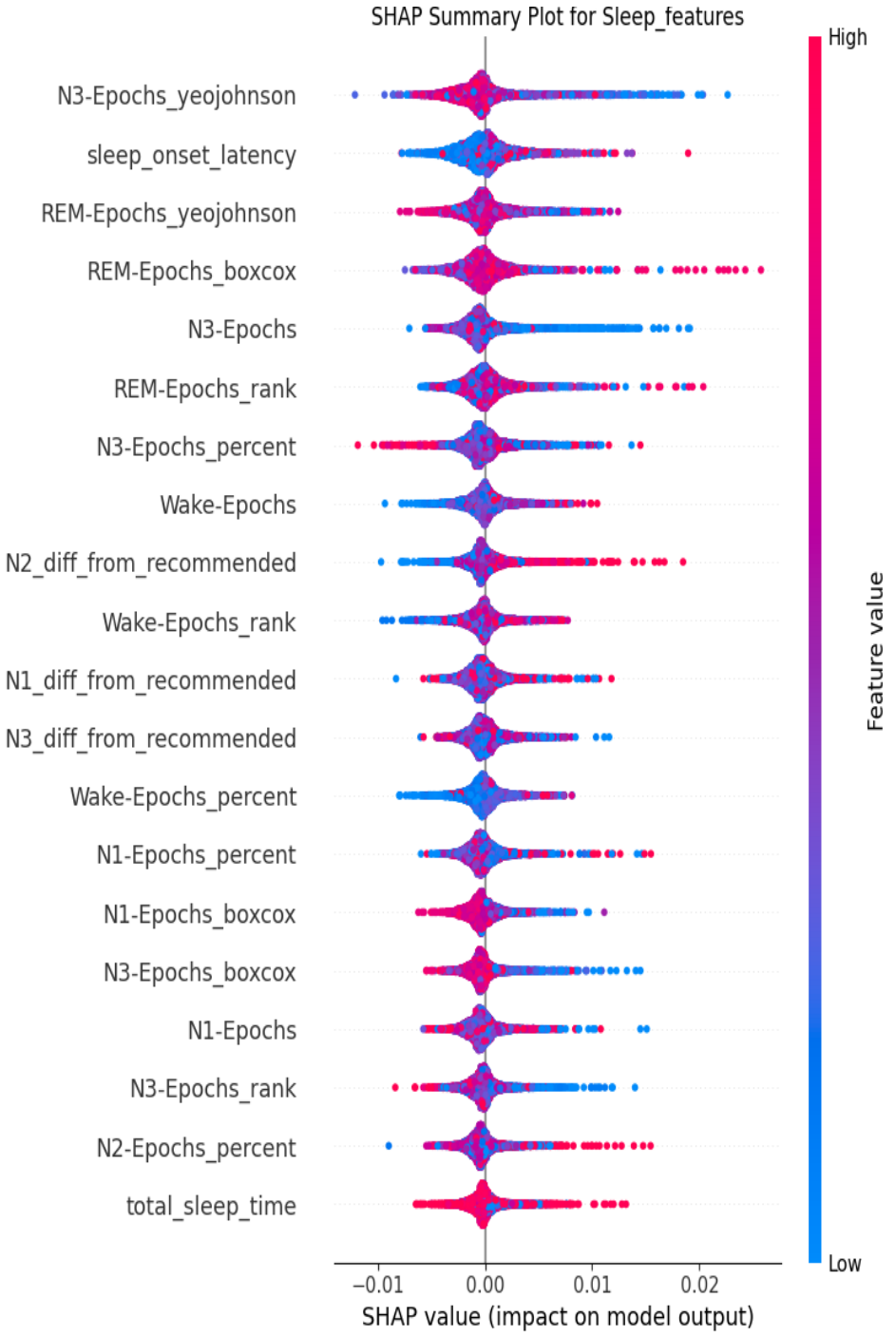
Features related to the time of day (cos_hour, Hourfrommid, Minfrommid) are among the most influential in the classification problem as well, underscoring the importance of circadian rhythms in glucose regulation. Physiological and Activity Metrics are also important for classification model, suggesting that the model captures the effects of stress, physical activity, and overall autonomic regulation on glucose classes. Various sleep metrics, particularly related to deep and REM sleep stages, are important for classification, indicating that sleep quality and structure may play a role in glucose fluctuations. The feature importances for classification for each data sources are plotted. The importance of sleep features is given in the picture below.

The SHAP visualizations are also plotted to gain insights into how different features impact the classification of IG classes into high, low, and normal classes. cos_hour representing the cosine transformation of the hour component, has the highest impact across all classes, especially high glucose (class 1) and normal glucose (Class 2). Hourfrommid and Minfrommid, representing time from midnight in hours and minutes, also show significant importance across classes. pNN50 is a heart rate variability (HRV) metric and quartiles (Q1G, Q3G) are particularly impactful, likely reflecting physiological responses that correlate with glucose classes. EDA and ACC features represent

electrodermal activity and accelerometer data, respectively, and they also contribute to the classification, indicating that physical activity and stress may play roles in glucose fluctuations.



(a)



(b)

Figure 5.24: SHAP explanations for Classification of IG values using an RF model. (a) Top 21 features using SHAP explanations for all classes. (b) SHAP values for sleep features used in prediction of IG classes using an RF model.

The SHAP values for various sleep-related features are also plotted. These values reveal that N3-Epochs (yeojohnson transformation), has a significant impact, particularly with higher values pushing the prediction towards high class. Sleep Onset Latency asleep has a moderate impact, with longer latencies slightly tilting the predictions. REM-Epochs Various transformations and rankings of REM sleep epochs contribute to the model, reflecting how REM sleep patterns may influence glucose regulation.

Features related to the time of day (cos_hour, Hourfrommid, Minfrommid) are among the most influential in the classification problem as well, underscoring the importance of circadian rhythms in glucose regulation. Physiological and Activity Metrics are also important for classification model, suggesting that the model captures the effects of stress, physical activity, and overall autonomic regulation on glucose classes. Various sleep metrics, particularly related to deep and REM sleep stages, are important for classification, indicating that sleep quality and structure may play a role in glucose fluctuations.

5.7 Discussion

Sleep features are divided into sleep parameters and sleep stages. These features and the methods to calculate these features are identified through an extensive literature review. These features are calculated using SleepPy by modifying the files to ensure compatibility with GeneActive data format which is what the SleepPy library can take as input. After that a public dataset that contains PSG labelled sleep is used to train an RF model to identify different stages in 30 second epochs of sleep which required the conversion of Empatica E4 data from glucose labelled data to Apple Watch format by aggregating the data and changing the values of accelerations to fractions of g. RFSLeeep model is trained and tested using a public dataset (Walch, 2019).

The predicted sleep stages are valuable inputs for a glucose prediction model. The overall patterns and transitions between sleep stages help the model learn body's physiological state in response to sleep, which significantly impacts glucose regulation. The integration of sleep stage information, alongside other features, provides a richer context for predicting glucose classes effectively.

5.7.1 Correlation of sleep features with IG labels

After finding sleep features, certain tests are carried out for normality and a correlation with IG values is taken with the help of Pearson and Spearman correlation. To improve the correlation of features with IG vales, the features are transformed using Log, Box-Cox, Yeo-Johnson, and Rank Based Inverse transforms. The selected transformations demonstrate a systematic approach to refining the input data, ensuring that it aligns with the assumptions of the analytical methods employed (Wang et al., 2019). This selection process enhances the performance of the regression models by reducing the RMSE of DT and RF models but has no significant effect on the classification model performance.

There is a mixture of positive and negative correlations of the features with IG values indicated in Figures 5.8 and 5.9, indicating that some features are directly associated with increases in IG, while others are inversely related. Notably, several features exhibit very strong correlations with IG, as evidenced by extremely low p-values (e.g., $p<10-30$),

which suggest these relationships are statistically significant. The positive correlation between total sleep time, percent sleep time, sleep onset latency, and interstitial glucose (IG) values suggests that as these sleep metrics increase, so do the IG levels. A longer sleep onset latency indicates difficulty in falling asleep, which could be linked to stress, anxiety, or other factors that dysregulate glucose metabolism. Adequate time spent in N2 sleep has been linked to improved glucose metabolism.

5.7.2 Effect of Sleep features on IG prediction ML models

5.7.2.1 Effects of sleep on regression models

The performance metrics for each model and feature set combination helps compare the models and the impact of incorporating different sets of sleep-related features. In regression models. AdaBoost using only the base features, the model performs poorly, with negative R^2 and Adjusted R^2 values, indicating that the model is not fitting the data well. The MAE (Mean Absolute Error) is relatively high, and the RMSE (Root Mean Squared Error) reflects significant prediction errors. Incorporating `all_sleep_features`, 'transformed_sleep_features`, or `unmodified_sleep_features` slightly improves the metrics, but the overall model fit remains poor (R^2 remains negative). However, there is a small reduction in error metrics (MAE and RMSE). DT using only the base features the model performs relatively well with a positive R^2 value (~0.576), indicating a decent fit to the data. The MAE is lower than AdaBoost, and the RMSE is significantly reduced. Introducing sleep features improves the model further, with `all_sleep_features` providing the best performance (R^2 ~0.71). This indicates that sleep features add valuable information that helps improve the prediction accuracy. The MAE and RMSE are both reduced, showing improved model precision. KNN using based features the model performs poorly with negative R^2 values, indicating poor data fitting. The MAE and RMSE are both high, reflecting poor prediction accuracy. Incorporating sleep features does not significantly improve the model's performance, and the R^2 values remain negative. The MAE and RMSE remain high, suggesting that KNN may not be suitable for this task. LassoCV performs with an R^2 close to zero, indicating a very poor fit using only base features. The error metrics are relatively high, showing the model's inability to capture the data patterns effectively. The inclusion of sleep features does not significantly impact the model's performance, and the R^2 remains close to zero, with similar error metrics across all feature sets. This suggests that LassoCV struggles with this data, and regularization might be too strong. In contrast, RF shows strong performance with a high R^2 (~0.815), indicating a good fit on base features. The MAE and RMSE are both low, demonstrating good prediction accuracy. Including `all_sleep_features` slightly improves the model's performance, further increasing the R^2 (~0.866) and reducing error metrics. This suggests that sleep features are valuable in improving prediction accuracy with RF.

XGBoost also performs well with a high R^2 (~0.769), indicating a good fit. The MAE and RMSE are relatively low, demonstrating good prediction performance. Including `all_sleep_features` further improves the model, increasing the R^2 (~0.812) and reducing the error metrics. This suggests that sleep features are valuable in improving prediction accuracy with XGBoost.

RMSE between DT and RF models is compared for two different feature sets: `other_features` and `other_features + all_sleep_features`. With `other_features`, the RMSE values range from around 14 to 15.5, with a median close to 15. Adding `all_sleep_features` to the feature set reduces the RMSE, with values ranging from approximately 11.5 to 14, and the median is around 12.5. The RMSE values are much lower compared to DT. With `other_features`, RMSE values range from about 8.5 to 9.5, with a median close to 9. Adding `all_sleep_features` further reduces RMSE, with values ranging from approximately 8 to 8.5, and the median is slightly above 8. The paired t-test compares the mean RMSE between the feature sets for each model. Between these RF models the t-test statistic is 22.5806 with a p-value < 0.000005, indicating a significant difference in RMSE when `all_sleep_features` are added. In the DT Model the t-test statistic is 5.5480 with a p-value of 0.0004, also showing a significant difference when `all_sleep_features` are included. Both DT and RF models exhibit a reduction in RMSE when sleep features are included (`other_features + all_sleep_features`), suggesting that these features contribute positively to the model's performance.

5.7.2.2 Effect of sleep on classification models

In classification models, the performance of various algorithms—AdaBoost, DT, KNN, Logistic Regression, RF, SVM, and XGBoost—was compared. AdaBoost showed minimal variation across feature sets, while DT saw notable improvement with unmodified_sleep_features, achieving an accuracy of 0.752. KNN and Logistic Regression did not benefit significantly from sleep features, with poor overall accuracy. RF and XGBoost showed the most improvement, with RF reaching 0.827 accuracy using all_sleep_features, and XGBoost reaching 0.823 with transformed_sleep_features. The inclusion of sleep features consistently improved the performance of tree-based models, while simpler models such as KNN and Logistic Regression saw little benefit.

The comparison of accuracies across tree models (RF, DT, and XGBoost) showed that DT had the lowest accuracy across all feature sets, while RF and XGBoost consistently achieved higher accuracy. Sleep features improved the accuracy of both RF and XGBoost, with all_sleep_features yielding the highest accuracy in both models. XGBoost demonstrated the highest performance, though its results were more sensitive to the feature set used, while RF exhibited more stability.

5.7.3 Comparison with earlier works

In this study, incorporating sleep features alongside traditional time, ANS, statistical, and food features has led to notable improvements in predicting interstitial glucose (IG) values and classes. For instance, RF classification models in the current study achieved an accuracy of 82.7% and DT for regression achieved an RMSE of RMSE = 6.59 ± 0.33 mg/dL, significantly outperforming previous models that did not consider sleep features.

Beyond improving model performance, sleep features increase the explainability of glucose predictions. Understanding how sleep quality and duration influence glucose levels can offer actionable insights for personalized interventions, encouraging lifestyle changes such as improving sleep hygiene. This makes the integration of sleep features not only beneficial for prediction accuracy but also crucial for facilitating meaningful changes in daily routines to better manage glucose levels.

Table 5.9 compares the performance of the models that use sleep features in this work with earlier works.

Table 5.10: Comparison of models that use sleep features with earlier works

Work	Model	Features	Performance
(Bent, Cho, Wittmann, et al., 2021b)	Decision Tree (DT)	Time, ANS, Statistical, Food	RMSE = 21.22 ± 4.14 mg/dL
(Zahedani et al., 2023)	LSTM	Time, ANS, Statistical, Food	RMSE = 14.8 mg/dL
(Ali et al., 2023)	XGBoost	Time, ANS, Statistical, Food	Accuracy = 64% to 86%
This work	Random Forest (RF)	Time, ANS, Statistical, Food, Sleep	Accuracy = 82.7%
This work	Decision Tree (DT)	Time, ANS, Statistical, Food, Sleep	RMSE = 6.59 ± 0.33 mg/dL

5.8 Conclusions

This study explored the intricate relationship between sleep features, physiological parameters, and interstitial glucose (IG) classes, providing valuable insights into the predictive power of these variables in glucose regulation. The findings demonstrate that sleep stages, particularly deep sleep (N2 and N3) and the duration of wake epochs, are

significantly correlated with IG classes, underscoring the critical role of sleep quality and duration in metabolic health.

The integration of sleep features into glucose prediction models, particularly with Random Forest (RF) and XGBoost, significantly enhanced predictive accuracy, as evidenced by lower RMSE values and improved R^2 metrics. These models effectively captured the complex relationships between sleep and glucose dynamics, outperforming others like K-Nearest Neighbors (KNN) and LassoCV, which struggled with the added complexity of sleep data. The analysis revealed that time-related features and Autonomic Nervous System (ANS) indicators were the most influential in glucose prediction, followed by sleep features, highlighting the multifaceted nature of glucose regulation.

The study's systematic approach, including the transformation of sleep data to reduce skewness and the rigorous evaluation of multiple machine learning models, reinforces the importance of selecting and preprocessing features that align with the assumptions of predictive models. The successful application of sleep features in enhancing model performance suggests that incorporating behavioural and physiological data from smartwatches could offer a more comprehensive approach to managing glucose variability, particularly for individuals with diabetes.

Beyond improving model performance, sleep features increase the explainability of glucose predictions. Understanding how sleep quality and duration influence glucose levels can offer actionable insights for personalized interventions, encouraging lifestyle changes such as improving sleep hygiene. This makes the integration of sleep features not only beneficial for prediction accuracy but also crucial for facilitating meaningful changes in daily routines to better manage glucose levels.

In conclusion, this research provides a robust framework for integrating sleep and physiological features into glucose prediction models, offering potential avenues for personalized interventions. Future work could focus on refining these models further by incorporating additional behavioural and environmental factors, exploring the temporal dynamics of glucose regulation, and validating these findings in larger, more diverse populations. This integrated approach to understanding glucose variability could lead to more effective strategies for managing metabolic health.

5.9 Chapter Summary

The findings presented in Chapter 5 highlight that the performance increase in ML models when sleep features are used. This chapter used open-source datasets to find sleep features from HR sensors and accelerometers. This Chapter answers research question 4 by first identifying sleep features predictive of IG values, correlating them with IG values and using them to illustrate the decrease in error term of IG prediction from ML

models. Tree based models see the biggest decrease in error term in comparison to other models in predicting glucose values and levels. Building upon these findings – a clinical decision support system can qualify the effect sleep has on an individual scale and provide related advice.

6 Discussion and Conclusion

The research questions that this thesis address is:

1. What is the state of the art of data processing and Machine Learning (ML) applications in time domain healthcare data?
2. What is the state of the art in digital biomarker design in predicting interstitial glucose (IG) from smart watch sensors?
3. How do different ML models compare in predicting IG from smart watch and food log data?
4. How does sleep parameters measured using HR and accelerometer data affect the performance of different ML models in predicting IG levels?

In view of this, this thesis is arranged into six Chapters. The first Chapter highlights the motivation for the thesis. Chapter 1 highlighted that a growing number of people with metabolic conditions across the world, including New Zealand. Whether it is a metabolic syndrome which is reversible set of risk factors or metabolic disease, a chronic condition, glucose levels monitoring is beneficial. Because both elevated and decreased levels of glucose can have lasting effect on the health such as insulin resistance and inability to recognize oncoming hypoglycaemia. This monitoring of glucose levels is typically done using CGM that measure interstitial glucose (IG) every one to five minutes or glycated haemoglobin tests (HbA1C) which measures two-to-three-month effect on metabolic health. These however do not measure the effect of activity or lifestyle changes on the glucose levels. Smartwatches are an increasingly prevalent technology, and they contain advanced sensors that can be used to measure various aspects of human health like Heart Rate Variability (HRV- a surrogate of stress) or accelerometer values (indicative of activity levels). These measurements have been shown to influence glucose levels in the body, hence can be used by ML models to predict IG. The benefit of these models is three folds, one it proves that ML models can be trained to design global glucose models given representative data and explanations of these models can help measure the state of health in a community, two, even if they do not capture a global model of glucose change based on smart watch data, they do contain information about how these two effect each other defined by the predictive performance of the ML models and three, models thus designed and explained using shapely additive explanations (SHAP) can inform users and clinicians to target lifestyle changes to manage metabolic health, lowering the risk of disease.

However, an investigation into design of such a ML model that uses time logged smart watch data merits a broader overview of how time series data is generally treated in the healthcare applications, this gives rise to the research question one, mentioned in the

beginning of this Chapter. This broader literature review is conducted in Chapter 2 of this thesis. Chapter 2 answers the question: *What is the state of the art of data processing and ML applications in time domain healthcare data?* It highlights different problems that can arise in developing ML applications from time domain electronic medical record (EMR) data, including levels of automation, privacy, interoperability of disparate devices, different kinds of decision systems. Chapter 2 also investigates the top methods used in these different aspects of model design and the problems faced. Chapter 2 also highlights the prospective future trends effectively giving the state of the art in data processing and ML applications in time domain healthcare data.

The broader literature review highlights the applications that use time series data and lists methods to overcome the issues faced, but a narrower literature review is needed to investigate how ML is being applied to predict the glucose levels from smart watch devices. These ties into research question one which reads: *What is the state of the art in digital biomarker design in predicting interstitial glucose (IG) from smart watch sensors?* Chapter 3 not only identifies the digital biomarkers calculated with smart watch data but also list practical steps to measure them. In Chapter 3, a systemic literature search is conducted with the help of a search term using relevant scientific databases. The records that are identified in this literature search are used to answer questions of the recently proposed Data, Aggregation, Contextualization, Integration and Action (DACIA) framework. DACIA framework proposes a standardised method to find digital biomarkers from raw data. This chapter lists the practical steps needed to convert the data into digital biomarkers and highlights the importance of different kinds of biomarkers in IG predictions from ML models. This review answers the research question 2 by investigating the state of the art in digital biomarker design using a systematic literature review and synthesizing the information from the identified literature using a standard framework.

The literature search in Chapter 3, highlighted two major gaps; one, there is a need to systematically compare the models used in IG prediction from smart watch sensors by keeping the feature set fixed and two, that sleep related biomarkers measured from smart watches are not used in studies that predict IG levels despite their reported effect on glucose regulation. These gaps are related to research questions three and four.

Chapter 4 answers the research question three by comparing ML models used to predict glucose levels from smart watch data and food logs. First, the smart watch (Empatica E4) data is filtered using relevant filtering techniques identified in Chapter 2 and 3. Then the data is converted into windows and digital biomarkers following the guidance from Chapter 3. The digital biomarkers are then used to classify IG values into personalized high, low and normal following the literature from Chapter 3. Finally different ML models

are trained to classify and predict IG values. These models are then compared using statistical tests and explained using Shapely Additive Explanations (SHAP). The performance of tree models has been shown to be robust to noise (in the presence of influential outliers shown in Cook's plot in Chapter 4). It also can measure complex interactions between chapter shown with the help of partial dependence plots.

As a result of the literature search in Chapter 3, sleep was identified as a missing biomarker type in IG prediction studies that use ML. This is related to the research question 4. Chapter 5 measures the effectiveness of sleep parameters measured using smart watch (Empatica E4) data in predicting glucose levels. Sleep biomarkers are divided into two categories sleep parameters and sleep stages. Sleep parameters include number of Wake Bouts (#WB) Wake After Sleep Onset (WASO), Sleep Onset Latency (SOL) and Total Sleep Time (TST). Whereas sleep stages measured every 30 seconds of sleep are non-rapid eye movement stages 1, 2 and 3 (N1, N2 and N3), Wake (W) and Rapid Eye Movement (REM). Sleep parameters are measured using rule-based methods whereas sleep stages are measured by training an ML model (RFSleep) on a labelled dataset. These parameters are used to define further novel sleep features. All of sleep features are correlated with the label values and the results corroborated with earlier findings. The new sleep features are added to the features identified in Chapter 3 and 4, and the increase in performance of different ML models is compared for both the classification and regression task, showing an increase in performance. This effectively answers the research question number 4.

A summary of research questions and how they are answered in this thesis are given in Table 6.1

Table 6.1: Summary of key findings with research questions

Chapter	Question	Main findings
2	What is the state of the art of data processing and ML applications in time domain healthcare data?	For time domain EMR data various paradigms are identified with the help of a systematic literature review. The research was classified into decision systems, preprocessing techniques, types of data, types of sensing elements, data preprocessing, levels of automation and explainability techniques. A combination of these methods can be used to perform various downstream tasks.
3	What is the state of the art in digital biomarker design in predicting interstitial glucose (IG) from smart watch sensors?	A systematic literature review was conducted to find different studies that use smart watch data to predict glucose. DACIA framework's guiding questions are used to find various practical steps needed to calculate the digital biomarkers. This research shows that the choice of digital biomarkers depends on the type of smart watch and the glucose markers to be predicted.
4	How do different ML models compare in predicting IG from smart watch and food log data?	A systematic comparison of different ML methods was carried out for both classification of IG levels into personalized high, low and normal glucose as well as prediction of IG values. Features identified in Chapter 3 are used to train the ML models. In both cases tree-based models (Random Forest (RF), Decision Tress (DT), and Extreme Gradient Boosting (XGBOOST) models outperform other models (Support Vector Machine (SVM), LASSO, Gaussian Naïve Bayes (GNB) and Linear Regression). It was shown that RF has the smallest mean absolute error (MAE) = 6.243 mg/dL and GNB has the highest MAE of 60.84 mg/dL. The performance of RF is explained using partial dependence plots and SHAP values which show RF models are better at modelling complex feature interactions and cook's plots show the influential outliers which is a possible explanation of why RF models outperform other models.
5	How does sleep parameters measured using HR and accelerometer data affect the performance of different ML models in predicting IG levels?	Sleep parameters are measured using accelerometer values using rule-based methods to measure number of wake bouts (#WB), Wake After Sleep Onset (WASO), Sleep Onset Latency (SOL) and Total Sleep Time (TST). ML model (RFSleep) was trained to predict the sleep stages (Non Rapid Eye Movement 1,2 and 3 (N1, N2 and N3) , Wake (W) and Rapid Eye Movement (REM). The features identified in Chapter 3 are used alongside sleep features derived from the sleep parameters and stages. Inclusion of sleep features increased the performance of RF model from MAE of 6.243 ± 0.121 mg/dL to 5.267 ± 0.133 mg/dL. These increases are verified using t-tests and Nemenyi post hoc analysis. SHAP value of sleep features for RF model shows that W and N2 epochs have very strong effect on the IG values predicted.

6.1 Novel Contributions

This thesis has following novel contributions in comparison to the earlier works in this field. Earlier works have utilized features developed using smart watch data to predict different glucose control markers, but no work has been done to systematically and empirically compare the models and the features. This work has the following novel contributions:

1- This thesis has compiled the relevant literature that utilizes time domain data in ML models and explains the journey of data to decision-making using EMR time domain data as an example.

2- This thesis has compiled the relevant literature that utilizes smart watch data to predict different glucose control markers using ML models. It focuses on practical steps to measure the digital biomarkers measured using smart watch data. The focus on listing pre-processing, windowing and filtering steps needed to calculate food, movement, circadian features, autonomic nervous system and physiology related features. This synthesis of the information is carried out using systematically. It also compares the relative importance of the biomarkers based on earlier work. This effectively allows for development of open-source libraries that can convert data into relevant biomarkers.

3- This thesis has compared different ML models, in comparison to earlier works, this work uses empirical data and statistical tests to compare the models and identifies the best performing models using statistical tests. This work also has explained the best performing models (RF models) using SHAP values, which is a novel contribution in comparison of earlier works, that have used relative importance of features. SHAP value explanations allow for which features have resulted in positive and negative correlations and the spread of those relationships giving a more nuanced understanding. Another novel feature of this work is the use of partial dependence plots that explains complex feature explanations.

4- This thesis employs sleep related features in prediction of glucose levels and uses statistical tests to prove the increase in performance of IG predicting ML models. The earlier works have not used sleep parameters or features to predict IG values. This contribution allows for advising targeted sleep related advice from clinicians in the context of metabolic disorders. The importance of sleep features is verified as well as SHAP values indicating how each sleep feature, affects the IG values as modelled by the best performing ML model,

6.2 Research Implications

Advances in in these Chapters can be leveraged by digital health experts and clinicians in monitoring more aspects of metabolic health that were inaccessible before.

By systematically compiling relevant literature on time-domain data used in ML models—particularly concerning EMR and smartwatch data—it offers a valuable reference for understanding the current state of research. This comprehensive collection aids in identifying knowledge gaps and provides a methodological framework that can be adopted or adapted for future studies.

Detailed guidelines for extracting digital biomarkers from smartwatch data are presented, focusing on practical steps like pre-processing, windowing, and filtering. This meticulous approach enables researchers to replicate processes accurately, ensuring consistency and reliability in data analysis. The potential development of open-source libraries based on this work further aids in standardizing methods for biomarker extraction, which is crucial for advancing research in this area.

The comparison of different ML models which help identify that tree-based models have lower error between actual and predicted IG values. This performance of the tree-based models informs, not just which models can serve as a baseline for future development also inform about the nature of the distribution of data and the feature interaction. The knowledge of distribution of data and feature interaction can be used to develop feature transformations that can help lower the error even further. The use of SHAP values and partial dependence plots also provides a nuanced understanding of feature importance and interactions, enhancing transparency and trust in machine learning applications within healthcare.

The sleep biomarkers descriptive of the IG values, developed in this work can be used by other deep learning models, as well inverse reinforcement learning models. Using inverse reinforcement learning techniques the human bodily systems can be treated as multi agent systems with some reward or value function that these systems are collaborating to optimise. The estimate of this reward function can inform the state of health and sleep inclusion as a state can increase the model's descriptive ability.

The effect of biomarkers on glucose levels modelled as SHAP values of the RF model predicting sleep values can be used by Large Language Models (LLM) to give personalized advice. LLMs in conjunction with healthcare ontologies can leverage both the quantitative data and the access to healthcare knowledge basis. ML models aim to minimize the error between numerical terms or learn representations of the data that aid in determining information patterns within the data.

6.3 Clinical Implications

By providing a comprehensive understanding of how time-domain data from electronic medical records and smartwatches can be utilized in machine learning models, clinicians gain clearer insights into data-driven decision-making processes. This transparency

enhances trust in predictive models, leading to more informed and confident clinical decisions.

The standardization of digital biomarker extraction from smartwatch data, focusing on practical steps like pre-processing, windowing, and filtering, ensures consistent and accurate measurement of patient health indicators. This consistency enhances the quality of patient monitoring by providing reliable data on physical activity, dietary intake, circadian rhythms, autonomic nervous system activity, and physiological parameters. As a result, clinicians can implement more precise interventions based on dependable data.

Understanding the strengths of the top performing ML models for IG prediction, supported by statistical tests, facilitates their confident integration into clinical practice. Additionally, the use of SHAP values provides deeper insights into how specific features influence IG predictions. This enables clinicians to comprehend the underlying factors affecting a patient's glucose levels and tailor interventions accordingly.

Incorporating sleep-related features into glucose prediction models acknowledges the significant impact of sleep on metabolic health. Clinicians can now consider sleep quality and patterns as essential factors in managing glucose levels, allowing for holistic treatment plans that include sleep hygiene recommendations. This approach potentially improves patient outcomes by addressing a previously underutilized aspect of metabolic control.

The models developed in this thesis can have the following pathways to be utilized in clinical settings. The amount of data generated by the longitudinal devices such as smartwatches is very high because of the high sampling frequency of the sensors, hence for these values to be used for health settings requires a condensed description of the effects of various aspects of data collected on metabolic health.

The models developed in this thesis and their respective explanations can be used by clinicians to infer which aspects of health (for example, increase in N2 sleep) have affected the IG values at an individual level. Each statistical property of the sensor signal signifies some aspect of physiology, for example, the expected value of accelerometer signals defined as mean of the signal defines how much the body has moved during the window of interest, and since how much someone moves is directly proportional to your energy expenditure, it is reasonable to assume that high values of mean accelerometer values would result in higher values of IG on average and this is evident in the SHAP values of the regression models in chapters 4 and 5. This descriptive ability of the models helps clinicians and users to reason about what causes an increase or decrease in their IG levels on a personal scale.

Utilizing non-invasive wearable technology for glucose monitoring empowers patients to actively engage in their health management. Access to immediate feedback on how

lifestyle choices affect glucose levels can motivate patients to adopt healthier behaviours. Enhanced patient engagement often leads to better adherence to treatment plans and can improve long-term health outcomes.

The rigorous statistical validation and methodological transparency of the IG prediction ML models support the development of data-driven clinical guidelines. Clinicians can rely on these evidence-based models to inform best practices, leading to standardized care protocols that improve patient outcomes across different healthcare settings.

Aggregated and anonymized data from wearable devices can contribute to public health surveillance, helping identify trends in metabolic health across populations. This information can inform public health initiatives and resource allocation, ultimately improving community health outcomes.

6.4 Limitations and future directions

This research aims to use ML models to predict glucose levels from smart watch sensors. As George Box famously said: “All models are wrong, but some are useful (Box, 1979)”. These models are not without errors. The error rate (the smallest error being on average 5 mg/dL) shows that these model’s prediction in its current form should not be taken as clinical advice. But with this caveat in mind, it is important to realize that these models are designed to prove that the values of glucose can be modelled with smart watches, as the error in glucose level classification models exceed chance levels. With further advances in model design and sensor technology, this error limit will be reduced, in fact, in the last three years it has reduced from 25 mg/dL to 5 mg/dL, thus giving the author hope that these advances are on the horizon.

Further biomarkers can be defined because of the vast body of ongoing research that models various aspects of physiology that affects glucose levels, with the help of non-invasive and self-updating smart watches. For example, recent works have used PPG sensors commonly found in smart watches to find continuous and cuff less blood pressure, this opens the avenue for using thus measured blood pressure as a digital biomarker predictive of glucose levels.

Additional biomarkers may be defined using novel research in sensor technologies such as ingestible sensors. There is a growing interest in research on sensors that measure brain activity, brain activity is shown to be an energy intense process. With advances non-intrusive monitoring of brain activity, it can be used to define novel digital biomarkers descriptive of glucose level changes. This may also broaden the understanding of how the brain functions and what processes and sub processes may result in higher expenditure of energy and increased glucose levels.

In future research, the biomarkers designed in this work, can be used to define events with a minimal description length so that clinicians can make sense of the vast amount

of data that is at their disposal in the form of numbers. The models involving these events may unearth patterns of changes that cause glucose level changes at an individual level, thus providing clinicians with the useful information needed to tailor their advice to the individual's needs and behavioural patterns (Mergenthaler et al., 2013).

In future research, deep learning models such as recurrent convolutional neural networks (RCNN) and long short term memory (LSTM) models can be used to learn the features of the sensor data directly and see if this results in a better performance,

In future research directions, additional features related to stress, diet variations, or medications should be engineered to further enhance the performance of the IG predicting ML models.

Digital twins of various human physiological systems aim to model the changes in the body, mathematically informed by real time data. Recent advances in glucose modelling using digital twins only rely on nutritional data to assess glucose changes in virtual diabetes patients (Cappon et al., 2023). These models can benefit with advances in digital biomarkers to ground their glucose level assessments in other aspects of human behaviour and physiology captured using smart watches.

The advances in LLM technology, and agentic systems can be trained on the representations of the data using the digital biomarkers, equipped with tools and having access to healthcare literature can serve as assistants to clinicians, and help in answering their questions in natural language. This can reduce the load on the clinicians to make sense of the trends within the changes in these digital biomarkers.

In Summary, this work opens avenues of research for utility of digital biomarkers in prediction of glucose levels and monitoring the metabolic health.

References

Abbas, H., Zahed, K., Alic, L., Zhu, Y., Sasangohar, F., Mehta, R., Lawley, M., Abbasi, Q. H., & Qaraqe, K. A. (2018). A wearable, low-cost hand tremor sensor for detecting hypoglycemic Events in diabetic patients. *2018 IEEE International RF and Microwave Conference (RFM)*, 182–184. https://ieeexplore.ieee.org/abstract/document/8846546/

Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., Denecke, K., & others. (2020). Technical metrics used to evaluate health care chatbots: Scoping review. *Journal of Medical Internet Research*, *22*(6), e18301–e18301.

Abedi, M., Hempel, L., Sadeghi, S., & Kirsten, T. (2022). GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Applied Sciences*, *12*(14), 7075.

Adam, T. C., & Epel, E. S. (2007). Stress, eating and the reward system. *Physiology & Behavior*, *91*(4), 449–458. https://doi.org/10.1016/j.physbeh.2007.04.011

Adams, D., & Nsugbe, E. (2021). Predictive Glucose Monitoring for People with Diabetes Using Wearable Sensors. *Engineering Proceedings*, *10*(1), Article 1. https://doi.org/10.3390/ecsa-8-11317

Adithan, C. (2017). *Principles of translational science in medicine: From bench to bedside* (2nd ed.). Indian J Med Res. 2017 Mar;145(3):408-9. doi: 10.4103/0971-5916.211685.

Adnan, K., Akbar, R., Khor, S. W., & Ali, A. B. A. (2020). Role and challenges of unstructured big data in healthcare. *Data Management, Analytics and Innovation*, 301-323-301–323.

Aggarwal, A., Garhwal, S., & Kumar, A. (2018). HEDEA: a Python tool for extracting and analysing semi-structured information from medical records. *Healthcare Informatics Research*, *24*(2), 148-153-148–153.

Aguilar, M., Bhuket, T., Torres, S., Liu, B., & Wong, R. J. (2015). Prevalence of the Metabolic Syndrome in the United States, 2003-2012. *JAMA*, *313*(19), 1973–1974. https://doi.org/10.1001/jama.2015.4260

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). *Interpretable machine learning in healthcare*. 559-560-559–560.

Ahmed, Z. (2022). Multi-omics strategies for personalized and predictive medicine: Past, current, and future translational opportunities. *Emerging Topics in Life Sciences*, *6*(2), 215–225.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Alberti, K. G. M. M., Eckel, R. H., Grundy, S. M., Zimmet, P. Z., Cleeman, J. I., Donato, K. A., Fruchart, J.-C., James, W. P. T., Loria, C. M., Smith, S. C., International Diabetes Federation Task Force on Epidemiology and Prevention, Hational Heart, Lung, and Blood Institute, American Heart Association, World Heart Federation, International Atherosclerosis Society, & International Association for the Study of Obesity. (2009). Harmonizing the metabolic syndrome: A joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation*, *120*(16), 1640–1645. https://doi.org/10.1161/CIRCULATIONAHA.109.192644

Aldeer, M., Javanmard, M., & Martin, R. P. (2018). A review of medication adherence monitoring technologies. *Applied System Innovation*, *1*(2), 14–14.

Alhaddad, A. Y., Aly, H., Gad, H., Al-Ali, A., Sadasivuni, K. K., Cabibihan, J.-J., & Malik, R. A. (2022). Sense and Learn: Recent Advances in Wearable Sensing and Machine Learning for Blood Glucose Monitoring and Trend-Detection. *Frontiers*

in Bioengineering and Biotechnology, *10*. https://www.frontiersin.org/articles/10.3389/fbioe.2022.876672

Ali, H., Madanain, S., White, D., Akhter, M. N., & Niazi, I. K. (2024). From wearable activity trackers to Interstitial Glucose: Data to Insight- A proposed scientific journey. *Proceedings of the 2024 Australasian Computer Science Week*, 61–64. https://doi.org/10.1145/3641142.3641154

Ali, H., Madanian, S., Malik, N., White, D., Russel, B. K., & Niazi, I. K. (2023). Prediction of Interstitial Glucose Levels Through Wearable Sensors Using Machine Learning: 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2023. *Proceedings of the 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2023*. https://doi.org/10.1109/CSDE59766.2023.10487681

Alrefaie, Z. (2014). Brief Assessment of Supine Heart Rate Variability in Normal Weight, Overweight, and Obese Females. *Annals of Noninvasive Electrocardiology*, *19*(3), 241–246. https://doi.org/10.1111/anec.12120

Amelia, R., Harahap, J., Zulham, Fujiati, I. I., & Wijaya, H. (2024). Educational Model and Prevention on Prediabetes: A Systematic Review. *Current Diabetes Reviews*, *20*(6), e101023221945. https://doi.org/10.2174/0115733998275518231006074504

Aminifar, A., Lamo, Y., Pun, K. I., & Rabbi, F. (2019). *A practical methodology for anonymization of structured health data*.

Anmella, G., Corponi, F., Li, B. M., Mas, A., Garriga, M., Sanabra, M., Pacchiarotti, I., Valentí, M., Grande, I., & Benabarre, A. (2024). Identifying digital biomarkers of illness activity and treatment response in bipolar disorder with a novel wearable device (TIMEBASE): Protocol for a pragmatic observational clinical study. *BJPsych Open*, *10*(5), e137.

Arleth, T., Andreassen, S., Orsini-Federici, M., Timi, A., & Benedetti, M. M. (2000). A model of glucose absorption from mixed meals. *Modelling and Control in Biomedical Systems 2000 (Including Biological Systems)*, 307–312.

Aryal, A., & Becerik-Gerber, B. (2019). A comparative study of predicting individual thermal sensation and satisfaction using wrist-worn temperature sensor, thermal camera and ambient temperature sensor. *Building and Environment*, *160*, 106223. https://doi.org/10.1016/j.buildenv.2019.106223

Ash, J. A., & Rapp, P. R. (2014). A quantitative neural network approach to understanding aging phenotypes. *Ageing Research Reviews*, *15*, 44-50-44–50.

Australian Food Composition Database. (n.d.). Retrieved July 20, 2024, from https://afcd.foodstandards.gov.au/

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64-59–64.

B. Bent, M. Henriquez, & J. P. Dunn. (2021). Cgmquantify: Python and R Software Packages for Comprehensive Analysis of Interstitial Glucose and Glycemic Variability from Continuous Glucose Monitor Data. *IEEE Open Journal of Engineering in Medicine and Biology*, *2*, 263–266. https://doi.org/10.1109/OJEMB.2021.3105816

Babineau, J. (2014). Product review: Covidence (systematic review software). *Journal of the Canadian Health Libraries Association/Journal de l'Association Des Bibliothèques de La Santé Du Canada*, *35*(2), 68–71.

Bai, J., Di, C., Xiao, L., Evenson, K. R., LaCroix, A. Z., Crainiceanu, C. M., & Buchner, D. M. (2016). An Activity Index for Raw Accelerometry Data and Its Comparison with Other Activity Metrics. *PLOS ONE*, *11*(8), e0160644. https://doi.org/10.1371/journal.pone.0160644

Balakrishna, S., & Thirumaran, M. (2020). Semantic interoperability in IoT and big data for health care: A collaborative approach. In *Handbook of Data Science Approaches for Biomedical Engineering* (pp. 185-220-185–220). Elsevier.

Balasubramanian, A., Wang, J., & Prabhakaran, B. (2016). Discovering multidimensional motifs in physiological signals for personalized healthcare. *IEEE Journal of Selected Topics in Signal Processing*, *10*(5), 832-841-832–841.

Banda, J. M., Seneviratne, M., Hernandez-Boussard, T., & Shah, N. H. (2018). Advances in electronic phenotyping: From rule-based definitions to machine learning models. *Annual Review of Biomedical Data Science*, *1*, 53-68-53–68.

Bartolome, A., & Prioleau, T. (2022). A computational framework for discovering digital biomarkers of glycemic control. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00656-z

Batra, S., & Sachdeva, S. (2016). Organizing standardized electronic healthcare records data for mining. *Health Policy and Technology*, *5*(3), 226-242-226–242.

Battelino, T., Danne, T., Bergenstal, R. M., Amiel, S. A., Beck, R., Biester, T., Bosi, E., Buckingham, B. A., Cefalu, W. T., & Close, K. L. (2019). Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, *42*(8), 1593–1603.

Baum, E. B. (1988). On the capabilities of multilayer perceptrons. *Journal of Complexity*, *4*(3), 193-215-193–215.

Bavan, L., Surmacz, K., Beard, D., Mellon, S., & Rees, J. (2019). Adherence monitoring of rehabilitation exercise with inertial sensors: A clinical validation study. *Gait & Posture*, *70*, 211-217-211–217.

Baxi, V., Edwards, R., Montalto, M., & Saha, S. (2022). Digital pathology and artificial intelligence in translational medicine and clinical practice. *Modern Pathology*, *35*(1), 23–32.

Baxter, S. L., & Lee, A. Y. (2021). Gaps in standards for integrating artificial intelligence technologies into ophthalmic practice. *Current Opinion in Ophthalmology*, *32*(5), 431-438-431–438.

Beck, R. W., Bergenstal, R. M., Riddlesworth, T. D., Kollman, C., Li, Z., Brown, A. S., & Close, K. L. (2019). Validation of Time in Range as an Outcome Measure for Diabetes Clinical Trials. *Diabetes Care*, *42*(3), 400–405. https://doi.org/10.2337/dc18-1444

Bellamy, D., Celi, L., & Beam, A. L. (2020). Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv Preprint arXiv:2010.01149*.

Benedetti, D., Olcese, U., Frumento, P., Bazzani, A., Bruno, S., d'Ascanio, P., Maestri, M., Bonanni, E., & Faraguna, U. (2021). Heart rate detection by Fitbit ChargeHR ™: A validation study versus portable polysomnography. *Journal of Sleep Research*, *30*(6), e13346. https://doi.org/10.1111/jsr.13346

Bennett, C. M., Guo, M., & Dharmage, S. C. (2007). HbA $_{1c}$ as a screening tool for detection of Type 2 diabetes: A systematic review. *Diabetic Medicine*, *24*(4), 333–343. https://doi.org/10.1111/j.1464-5491.2007.02106.x

Bent, B., Cho, P. J., Henriquez, M., Wittmann, A., Thacker, C., Feinglos, M., Crowley, M. J., & Dunn, J. P. (2021). Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. *Npj Digital Medicine*, *4*(1), 89. https://doi.org/10.1038/s41746-021-00465-w

Bent, B., Cho, P. J., Wittmann, A., Thacker, C., Muppidi, S., Snyder, M., Crowley, M. J., Feinglos, M., & Dunn, J. P. (2021a). Non-invasive wearables for remote monitoring of HbA1c and glucose variability: Proof of concept. *BMJ Open Diabetes Research & Care*, *9*(1), e002027. https://doi.org/10.1136/bmjdrc-2020-002027

Bent, B., Cho, P. J., Wittmann, A., Thacker, C., Muppidi, S., Snyder, M., Crowley, M. J., Feinglos, M., & Dunn, J. P. (2021b). Non-invasive wearables for remote monitoring of HbA1c and glucose variability: Proof of concept. *BMJ Open Diabetes Research and Care*, *9*(1), e002027.

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *Npj Digital Medicine*, *3*(1), 1–9. https://doi.org/10.1038/s41746-020-0226-6

Bernardini, M., Romeo, L., Misericordia, P., & Frontoni, E. (2019). Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. *IEEE Journal of Biomedical and Health Informatics*, *24*(1), 235-246-235–246.

Bertachi, A., Viñals, C., Biagi, L., Contreras, I., Vehí, J., Conget, I., & Giménez, M. (2020). Prediction of Nocturnal Hypoglycemia in Adults with Type 1 Diabetes

under Multiple Daily Injections Using Continuous Glucose Monitoring and Physical Activity Monitor. *Sensors (Basel, Switzerland)*, *20*(6), 1705. https://doi.org/10.3390/s20061705

Bi, Z., Han, Y., Huang, C., & Wang, M. (2019). Gaussian naive Bayesian data classification model based on clustering algorithm. *2019 International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)*, 396–400. https://www.atlantis-press.com/proceedings/masta-19/125913250

Bialasiewicz, P., Pawlowski, M., Nowak, D., Loba, J., & Czupryniak, L. (2009). Decreasing concentration of interstitial glucose in REM sleep in subjects with normal glucose tolerance. *Diabetic Medicine: A Journal of the British Diabetic Association*, *26*(4), 339–344. https://doi.org/10.1111/j.1464-5491.2009.02684.x

Big Data Donation Project | *Tidepool*. (n.d.). Retrieved January 27, 2024, from https://www.tidepool.org/bigdata

Bizzego, A., Battisti, A., Gabrieli, G., Esposito, G., & Furlanello, C. (2019). pyphysio: A physiological signal processing library for data science approaches in physiology. *SoftwareX*, *10*, 100287. https://doi.org/10.1016/j.softx.2019.100287

Bjarnadottir, R. I., & Lucero, R. J. (2018). What can we learn about fall risk factors from EHR nursing notes? A text mining study. *eGEMs*, *6*(1).

Bohn, K., Amberg, M., Meier, T., Forner, F., Stangl, G. I., & Mäder, P. (2022). Estimating food ingredient compositions based on mandatory product labeling. *Journal of Food Composition and Analysis*, *110*, 104508. https://doi.org/10.1016/j.jfca.2022.104508

Bota, P., Silva, R., Carreiras, C., Fred, A., & da Silva, H. P. (2024). BioSPPy: A Python toolbox for physiological signal processing. *SoftwareX*, *26*, 101712. https://doi.org/10.1016/j.softx.2024.101712

Boudreau, P., Yeh, W.-H., Dumont, G. A., & Boivin, D. B. (2013). Circadian Variation of Heart Rate Variability Across Sleep Stages. *Sleep*, *36*(12), 1919–1928. https://doi.org/10.5665/sleep.3230

Boustani, M., Perkins, A. J., Khandker, R. K., Duong, S., Dexter, P. R., Lipton, R., Black, C. M., Chandrasekaran, V., Solid, C. A., & Monahan, P. (2020). Passive digital signature for early identification of Alzheimer's disease and related dementia. *Journal of the American Geriatrics Society*, *68*(3), 511-518-511–518.

Box, G. E. (1979). All models are wrong, but some are useful. *Robustness in Statistics*, *202*(1979), 549.

Boxwala, A. A., Kim, J., Grillo, J. M., & Ohno-Machado, L. (2011). Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association*, *18*(4), 498-505-498–505.

Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., & Huang, T. S. (2016). *Multi-modal audio, video and physiological sensor learning for continuous emotion prediction*. 97-104-97–104.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Briançon-Marjollet, A., Weiszenstein, M., Henri, M., Thomas, A., Godin-Ribuot, D., & Polak, J. (2015). The impact of sleep disorders on glucose metabolism: Endocrine and molecular mechanisms. *Diabetology & Metabolic Syndrome*, *7*(1), 25. https://doi.org/10.1186/s13098-015-0018-3

Brunner, C., & Hofer, F. (2023). SleepECG: A Python package for sleep staging based onheart rate. *Journal of Open Source Software*, *8*(86), 5411. https://doi.org/10.21105/joss.05411

Byun, J.-I., Cha, K. S., Jun, J. E., Kim, T.-J., Jung, K.-Y., Jeong, I.-K., & Shin, W. C. (2020). Dynamic changes in nocturnal blood glucose levels are associated with sleep-related features in patients with obstructive sleep apnea. *Scientific Reports*, *10*(1), 17877. https://doi.org/10.1038/s41598-020-74908-x

Cai, H., Qu, Z., Li, Z., Zhang, Y., Hu, X., & Hu, B. (2020). Feature-level fusion approaches based on multimodal EEG data for depression recognition. *Information Fusion*, *59*, 127-138-127–138.

Campbell, J. I., Eyal, N., Musiimenta, A., & Haberer, J. E. (2016). Ethical questions in medical electronic adherence monitoring. *Journal of General Internal Medicine*, *31*(3), 338-342-338–342.

Cappon, G., Facchinetti, A., Sparacino, G., Georgiou, P., & Herrero, P. (2019). Classification of Postprandial Glycemic Status with Application to Insulin Dosing in Type 1 Diabetes-An In Silico Proof-of-Concept. *Sensors (Basel, Switzerland)*, *19*(14), 3168. https://doi.org/10.3390/s19143168

Cappon, G., Vettoretti, M., Sparacino, G., Favero, S. D., & Facchinetti, A. (2023). ReplayBG: A Digital Twin-Based Methodology to Identify a Personalized Model From Type 1 Diabetes Data and Simulate Glucose Concentrations to Assess Alternative Therapies. *IEEE Transactions on Biomedical Engineering*, *70*(11), 3227–3238. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/TBME.2023.3286856

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., & Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics*, *4*(4), 195–202. https://doi.org/10.15406/ijbsbe.2018.04.00125

CDC. (2024a, May 21). *Prediabetes – Your Chance to Prevent Type 2 Diabetes*. Diabetes. https://www.cdc.gov/diabetes/prevention-type-2/prediabetes-prevent-type-2.html

CDC. (2024b, May 28). *National Diabetes Statistics Report*. Diabetes. https://www.cdc.gov/diabetes/php/data-research/index.html

Ceriello, A. (2005). Postprandial hyperglycemia and diabetes complications: Is it time to treat? *Diabetes*, *54*(1), 1–7.

Chai, P. R., Castillo-Mancilla, J., Buffkin, E., Darling, C., Rosen, R. K., Horvath, K. J., Boudreaux, E. D., Robbins, G. K., Hibberd, P. L., & Boyer, E. W. (2015). Utilizing an ingestible biosensor to assess real-time medication adherence. *Journal of Medical Toxicology*, *11*(4), 439-444-439–444.

Chai, P. R., Goodman, G., Bustamante, M., Mendez, L., Mohamed, Y., Mayer, K. H., Boyer, E. W., Rosen, R. K., & O'Cleirigh, C. (2021). Design and delivery of real-time adherence data to men who have sex with men using antiretroviral pre-exposure prophylaxis via an ingestible electronic sensor. *AIDS and Behavior*, *25*(6), 1661-1674-1661–1674.

Chandra, V., Priyarup, A., & Sethia, D. (2021). Comparative Study of Physiological Signals from Empatica E4 Wristband for Stress Classification. In M. Singh, V. Tyagi, P. K. Gupta, J. Flusser, T. Ören, & V. R. Sonawane (Eds.), *Advances in Computing and Data Sciences* (pp. 218–229). Springer International Publishing. https://doi.org/10.1007/978-3-030-88244-0_21

Chase, J. D., Busa, M. A., Staudenmayer, J. W., & Sirard, J. R. (2022). Sleep Measurement UsingWrist-Worn Accelerometer Data Compared with Polysomnography. *Sensors*, *22*(13). Scopus. https://doi.org/10.3390/s22135041

Chen, C., Jafari, R., & Kehtarnavaz, N. (2015). A real-time human action recognition system using depth and inertial sensor fusion. *IEEE Sensors Journal*, *16*(3), 773-781-773–781.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2020). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, *4*.

Chen, J. X. (2016). The evolution of computing: AlphaGo. *Computing in Science & Engineering*, *18*(4), 4-7-4–7.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*, 8869-8879-8869–8879.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, X., & Sun, L. (2022). Bayesian Temporal Factorization for Multidimensional Time Series Prediction. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, *44*(9), 4659–4673. Scopus. https://doi.org/10.1109/TPAMI.2021.3066551

Chin-Cheong, K., Sutter, T., & Vogt, J. E. (2019). *Generation of heterogeneous synthetic electronic health records using GANs*. workshop on machine learning for health (ML4H) at the 33rd conference on neural information processing systems (NeurIPS 2019).

Cho, P., Kim, J., Bent, B., & Dunn, J. (2023a). *BIG IDEAs Lab Glycemic Variability and Wearable Device Data*. https://doi.org/10.13026/zthx-5212

Cho, P., Kim, J., Bent, B., & Dunn, J. (2023b). *BIG IDEAs Lab Glycemic Variability and Wearable Device Data*.

Choi, J. M., Ji, M., Watson, L. T., & Zhang, L. (2023). DeepMicroGen: A generative adversarial network-based method for longitudinal microbiome data imputation. *Bioinformatics*, *39*(5). Scopus. https://doi.org/10.1093/bioinformatics/btad286

Christakis, Y. (2024). *Elyiorgos/sleeppy* [Python]. https://github.com/elyiorgos/sleeppy (Original work published 2019)

Colberg, S. R., Sigal, R. J., Fernhall, B., Regensteiner, J. G., Blissmer, B. J., Rubin, R. R., Chasan-Taber, L., Albright, A. L., Braun, B., American College of Sports Medicine, & American Diabetes Association. (2010). Exercise and type 2 diabetes: The American College of Sports Medicine and the American Diabetes Association: joint position statement. *Diabetes Care*, *33*(12), e147-167. https://doi.org/10.2337/dc10-9990

Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J., & Gillin, J. C. (1992). Automatic sleep/wake identification from wrist activity. *Sleep*, *15*(5), 461–469. https://doi.org/10.1093/sleep/15.5.461

Cooper, G. F., Aliferis, C. F., Ambrosino, R., Aronis, J., Buchanan, B. G., Caruana, R., Fine, M. J., Glymour, C., Gordon, G., Hanusa, B. H., & others. (1997). An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, *9*(2), 107-138-107–138.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. IEEE Transactions on Information Theory. https://doi.org/10.1109/TIT.1967.1053964

Coyle-Asbil, H. J., Habegger, J., Oliver, M., & Vallis, L. A. (2023). Enabling the ActiGraph GT9X Link's Idle Sleep Mode and Inertial Measurement Unit Settings Directly Impacts Data Acquisition. *Sensors*, *23*(12). Scopus. https://doi.org/10.3390/s23125558

Crofts, C., Schofield, G., Zinn, C., Wheldon, M., & Kraft, J. (2016). Identifying hyperinsulinaemia in the absence of impaired glucose tolerance: An examination of the Kraft database. *Diabetes Research and Clinical Practice*, *118*, 50–57.

Daly, B. M., Wu, Z., Nirantharakumar, K., Chepulis, L., Rowan, J. A., & Scragg, R. K. R. (2024). Increased risk of cardiovascular and renal disease, and diabetes for all women diagnosed with gestational diabetes mellitus in New Zealand—A national retrospective cohort study. *Journal of Diabetes*, *16*(4), e13535. https://doi.org/10.1111/1753-0407.13535

Daniore, P., Nittas, V., Haag, C., Bernard, J., Gonzenbach, R., & von Wyl, V. (2024). From wearable sensor data to digital biomarker development: Ten lessons learned and a framework proposal. *Npj Digital Medicine*, *7*(1), 1–8. https://doi.org/10.1038/s41746-024-01151-3

Dautov, R., Distefano, S., & Buyya, R. (2019). Hierarchical data fusion for Smart Healthcare. *Journal of Big Data*, *6*(1), 1-23-1–23.

Davy, C., Bleasel, J., Liu, H., Tchan, M., Ponniah, S., & Brown, A. (2015). Effectiveness of chronic care models: Opportunities for improving healthcare practice and health outcomes: A systematic review. *BMC Health Services Research*, *15*(1), 1-11-1–11.

de Looff, P., Duursma, R., Noordzij, M., Taylor, S., Jaques, N., Scheepers, F., de Schepper, K., & Koldijk, S. (2022). Wearables: An R Package With Accompanying Shiny Application for Signal Analysis of a Wearable Device Targeted at Clinicians and Researchers. *Frontiers in Behavioral Neuroscience*, *16*. https://doi.org/10.3389/fnbeh.2022.856544

Debnath, S., Levy, T. J., Bellehsen, M., Schwartz, R. M., Barnaby, D. P., Zanos, S., Volpe, B. T., & Zanos, T. P. (2021). A method to quantify autonomic nervous system function in healthy, able-bodied individuals. *Bioelectronic Medicine*, *7*(1), 13. https://doi.org/10.1186/s42234-021-00075-7

Dehghani Zahedani, A., Shariat Torbaghan, S., Rahili, S., Karlin, K., Scilley, D., Thakkar, R., Saberi, M., Hashemi, N., Perelman, D., Aghaeepour, N., McLaughlin, T., & Snyder, M. P. (2021). Improvement in Glucose Regulation Using a Digital Tracker and Continuous Glucose Monitoring in Healthy Adults and Those with Type 2 Diabetes. *Diabetes Therapy*, *12*(7), 1871–1886. https://doi.org/10.1007/s13300-021-01081-3

Dewettinck, K., Van Bockstaele, F., Kühne, B., Van de Walle, D., Courtens, T. M., & Gellynck, X. (2008). Nutritional value of bread: Influence of processing, food interaction and consumer perception. *Journal of Cereal Science*, *48*(2), 243–257. https://doi.org/10.1016/j.jcs.2008.01.003

Djenouri, D., & Balasingham, I. (2009). *New QoS and geographical routing in wireless biomedical sensor networks*. 1-8-1–8.

Dubosson, F., Ranvier, J.-E., Bromuri, S., Calbimonte, J.-P., Ruiz, J., & Schumacher, M. (2018). The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. *Informatics in Medicine Unlocked*, *13*, 92–100. https://doi.org/10.1016/j.imu.2018.09.003

Dugani, S. B., Lahr, B. D., Xie, H., Mielke, M. M., Bailey, K. R., & Vella, A. (2024). County Rurality and Incidence and Prevalence of Diagnosed Diabetes in the United States. *Mayo Clinic Proceedings*. https://www.sciencedirect.com/science/article/pii/S0025619623005724

Eades, C. E., Burrows, K. A., Andreeva, R., Stansfield, D. R., & Evans, J. Mm. (2024). Prevalence of gestational diabetes in the United States and Canada: A systematic review and meta-analysis. *BMC Pregnancy and Childbirth*, *24*(1), 204. https://doi.org/10.1186/s12884-024-06378-2

Ellebrecht, D. B., Gola, D., & Kaschwich, M. (2022). Evaluation of a Wearable in-Ear Sensor for Temperature and Heart Rate Monitoring: A Pilot Study. *Journal of Medical Systems*, *46*(12), 91. https://doi.org/10.1007/s10916-022-01872-6

Ervin, R. B. (2009). Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006. *National Health Statistics Reports*, *13*, 1–7.

Ewusie, J. E., Soobiah, C., Blondal, E., Beyene, J., Thabane, L., & Hamid, J. S. (2020). Methods, applications and challenges in the analysis of interrupted time series data: A scoping review. *Journal of Multidisciplinary Healthcare*, *13*, 411–411.

Finlayson, S. G., LePendu, P., & Shah, N. H. (2014). Building the graph of medicine from millions of clinical narratives. *Scientific Data*, *1*(1), 1-9-1–9.

Föll, S., Maritsch, M., Spinola, F., Mishra, V., Barata, F., Kowatsch, T., Fleisch, E., & Wortmann, F. (2021). FLIRT: A feature generation toolkit for wearable data. *Computer Methods and Programs in Biomedicine*, *212*, 106461. https://doi.org/10.1016/j.cmpb.2021.106461

Fonda, S. J., Graham, C., Munakata, J., Powers, J. M., Price, D., & Vigersky, R. A. (2016). The cost-effectiveness of real-time continuous glucose monitoring (RT-CGM) in type 2 diabetes. *Journal of Diabetes Science and Technology*, *10*(4), 898–904.

Ford, E. S., Li, C., & Sattar, N. (2008). Metabolic syndrome and incident diabetes: Current state of the evidence. *Diabetes Care*, *31*(9), 1898–1904.

Fujimoto, T., Nakajima, H., Tsuchiya, N., Marukawa, H., Kuramoto, K., Kobashi, S., & Hata, Y. (2013). Wearable Human Activity Recognition by Electrocardiograph and Accelerometer. *2013 IEEE 43rd International Symposium on Multiple-Valued Logic*, 12–17. https://doi.org/10.1109/ISMVL.2013.60

Gabrieli, G., Azhari, A., & Esposito, G. (2020). PySiology: A Python Package for Physiological Feature Extraction. In A. Esposito, M. Faundez-Zanuy, F. C. Morabito, & E. Pasero (Eds.), *Neural Approaches to Dynamics of Signal Exchanges* (pp. 395–402). Springer. https://doi.org/10.1007/978-981-13-8950-4_35

Ganie, S. M., Pramanik, P. K. D., Bashir Malik, M., Mallik, S., & Qin, H. (2023). An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics*, *14*. https://www.frontiersin.org/articles/10.3389/fgene.2023.1252159

Geraci, J., Wilansky, P., de Luca, V., Roy, A., Kennedy, J. L., & Strauss, J. (2017). Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evidence-Based Mental Health*, *20*(3), 83-87-83–87.

Goh, K. H., Wang, L., Yeow, A. Y. K., Poh, H., Li, K., Yeow, J. J. L., & Tan, G. Y. H. (2021). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, *12*(1), 1-10-1–10.

Golovaty, I., Ritchie, N. D., Tuomilehto, J., Mohan, V., Ali, M. K., Gregg, E. W., Bergman, M., & Moin, T. (2023). Two decades of diabetes prevention efforts: A call to innovate and revitalize our approach to lifestyle change. *Diabetes Research and Clinical Practice*, *198*, 110195.

Golshahi, J., Ahmadzadeh, H., Sadeghi, M., Mohammadifard, N., & Pourmoghaddas, A. (2015). Effect of self-care education on lifestyle modification, medication adherence and blood pressure in hypertensive adults: Randomized controlled clinical trial. *Advanced Biomedical Research*, *4*.

Gomes, P. (2022). *PyHRV: Python Toolbox for Heart Rate Variability*.

Gopinath, D., Agrawal, M., Murray, L., Horng, S., Karger, D., & Sontag, D. (2020). *Fast, Structured Clinical Documentation via Contextual Autocomplete* (F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, & J. Wiens, Eds.; Vol.

126, pp. 842-870-842–870). PMLR. http://proceedings.mlr.press/v126/gopinath20a.html

Gordon, W. J., & Catalini, C. (2018). Blockchain technology for healthcare: Facilitating the transition to patient-driven interoperability. *Computational and Structural Biotechnology Journal*, *16*, 224-230-224–230.

Grandner, M. A., Bromberg, Z., Hadley, A., Morrell, Z., Graf, A., Hutchison, S., & Freckleton, D. (2023). Performance of a multisensor smart ring to evaluate sleep: In-lab and home-based evaluation of generalized and personalized algorithms. *Sleep*, *46*(1), zsac152. https://doi.org/10.1093/sleep/zsac152

Grundy, S. M., Brewer, H. B., Cleeman, J. I., Smith, S. C., Lenfant, C., American Heart Association, & National Heart, Lung, and Blood Institute. (2004). Definition of metabolic syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation*, *109*(3), 433–438. https://doi.org/10.1161/01.CIR.0000111245.75752.C6

Gu, D., Li, T., Wang, X., Yang, X., & Yu, Z. (2019). Visualizing the intellectual structure and evolution of electronic health and telemedicine research. *International Journal of Medical Informatics*, *130*, 103947–103947.

Gu, W., Zhou, Y., Zhou, Z., Liu, X., Zou, H., Zhang, P., Spanos, C. J., & Zhang, L. (2017). SugarMate: Non-intrusive Blood Glucose Monitoring with Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 54:1-54:27. https://doi.org/10.1145/3130919

Guan, Z., Lv, Z., Du, X., Wu, L., & Guizani, M. (2019). Achieving data utility-privacy tradeoff in Internet of medical things: A machine learning approach. *Future Generation Computer Systems*, *98*, 60-68-60–68.

Gupta, Y., Goyal, A., Ambekar, S., Kalaivani, M., Bhatla, N., & Tandon, N. (2024). Cardiometabolic profile of women with a history of overt diabetes compared to gestational diabetes and normoglycemia in index pregnancy: Results from

CHIP-F study. *Journal of Diabetes*, *16*(5), e13461. https://doi.org/10.1111/1753-0407.13461

Ha, G.-B., Steinberg, B. A., Freedman, R., Bayés-Genís, A., & Sanchez, B. (2023). Safety evaluation of smart scales, smart watches, and smart rings with bioimpedance technology shows evidence of potential interference in cardiac implantable electronic devices. *Heart Rhythm*, *20*(4), 561–571.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., & Smith, N. J. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Hartl, D., de Luca, V., Kostikova, A., Laramie, J., Kennedy, S., Ferrero, E., Siegel, R., Fink, M., Ahmed, S., Millholland, J., & others. (2021). Translational precision medicine: An industry perspective. *Journal of Translational Medicine*, *19*(1), 1-14-1–14.

Hartono, P. (2020). A transparent cancer classifier. *Health Informatics Journal*, *26*(1), 190-204-190–204.

Harvey, C. J. d. C., Schofield, G. M., Zinn, C., Thornley, S. J., Crofts, C., & Merien, F. L. R. (2019). Low-carbohydrate diets differing in carbohydrate restriction improve cardiometabolic and anthropometric markers in healthy adults: A randomised clinical trial. *PeerJ*, *7*, e6273. https://doi.org/10.7717/peerj.6273

HAYERI, A. (2018). Predicting Future Glucose Fluctuations Using Machine Learning and Wearable Sensor Data. *Diabetes*, *67*(Supplement_1), 738-P. https://doi.org/10.2337/db18-738-P

Hees, V. T. van, Sabia, S., Anderson, K. N., Denton, S. J., Oliver, J., Catt, M., Abell, J. G., Kivimäki, M., Trenell, M. I., & Singh-Manoux, A. (2015). A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. *PLOS ONE*, *10*(11), e0142533. https://doi.org/10.1371/journal.pone.0142533

Heller, S. R., Buse, J. B., Ratner, R., Seaquist, E., Bardtrum, L., Hansen, C. T., Tutkunkardas, D., & Moses, A. C. (2019). Redefining Hypoglycemia in Clinical Trials: Validation of Definitions Recently Adopted by the American Diabetes

Association/European Association for the Study of Diabetes. *Diabetes Care*, *43*(2), 398–404. https://doi.org/10.2337/dc18-2361

Herzog, N., Jauch-Chara, K., Hyzy, F., Richter, A., Friedrich, A., Benedict, C., & Oltmanns, K. M. (2013). Selective slow wave sleep but not rapid eye movement sleep suppression impairs morning glucose tolerance in healthy men. *Psychoneuroendocrinology*, *38*(10), 2075–2082. https://doi.org/10.1016/j.psyneuen.2013.03.018

Higgins, M. K. (2021). Can we AlphaFold our way out of the next pandemic? *Journal of Molecular Biology*, 167093–167093.

Home, Resources, diabetes, L. with, Acknowledgement, FAQs, Contact, & Policy, P. (2021). *IDF Diabetes Atlas*. https://diabetesatlas.org/

Hong, N., Wen, A., Stone, D. J., Tsuji, S., Kingsbury, P. R., Rasmussen, L. V., Pacheco, J. A., Adekkanattu, P., Wang, F., Luo, Y., & others. (2019). Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *Journal of Biomedical Informatics*, *99*, 103310–103310.

Hossain, M. S., & Muhammad, G. (2017). Emotion-aware connected healthcare big data towards 5G. *IEEE Internet of Things Journal*, *5*(4), 2399-2406-2399–2406.

Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G., Liu, S., Solomon, C. G., & Willett, W. C. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *The New England Journal of Medicine*, *345*(11), 790–797. https://doi.org/10.1056/NEJMoa010492

Hu, J., Fang, M., Pike, J. R., Lutsey, P. L., Sharrett, A. R., Wagenknecht, L. E., Hughes, T. M., Seegmiller, J. C., Gottesman, R. F., & Mosley, T. H. (2023). Prediabetes, intervening diabetes and subsequent risk of dementia: The Atherosclerosis Risk in Communities (ARIC) study. *Diabetologia*, *66*(8), 1442–1449.

Huang, X., Schmelter, F., Seitzer, C., Martensen, L., Otzen, H., Piet, A., Witt, O., Schröder, T., Günther, U., Grzegorzek, M., & Sina, C. (2023). *From Data to Insight: Predicting Interstitial Glucose in Healthy Cohort with Non-invasive*

Sensor Technology and Machine Learning. https://doi.org/10.21203/rs.3.rs-3008236/v1

Hutchison, A. T., Regmi, P., Manoogian, E. N. C., Fleischer, J. G., Wittert, G. A., Panda, S., & Heilbronn, L. K. (2019). Time-Restricted Feeding Improves Glucose Tolerance in Men at Risk for Type 2 Diabetes: A Randomized Crossover Trial. *Obesity*, *27*(5), 724–732. https://doi.org/10.1002/oby.22449

Jabbar, R., Fetais, N., Krichen, M., & Barkaoui, K. (2020). *Blockchain technology for healthcare: Enhancing shared electronic health record interoperability and integrity*. 310-317-310–317.

Jackson, M. A., Ahmann, A., & Shah, V. N. (2021). Type 2 Diabetes and the Use of Real-Time Continuous Glucose Monitoring. *Diabetes Technology & Therapeutics*, *23*(Suppl 1), S-27-S-34. https://doi.org/10.1089/dia.2021.0007

Jahromi, R., Zahed, K., Sasangohar, F., Erraguntla, M., Mehta, R., & Qaraqe, K. (2023). Hypoglycemia Detection Using Hand Tremors: Home Study of Patients With Type 1 Diabetes. *JMIR Diabetes*, *8*(1), e40990. https://doi.org/10.2196/40990

Jain, S. (2024). 'Prediabetes' as a practical distinctive window for workable fruitful wonders: Prevention and progression alert as advanced professionalism. *World Journal of Clinical Pediatrics*, *13*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11000058/

Jarvis, P. R. E., Cardin, J. L., Nisevich-Bede, P. M., & McCarter, J. P. (2023). Continuous glucose monitoring in a healthy population: Understanding the post-prandial glycemic response in individuals without diabetes mellitus. *Metabolism - Clinical and Experimental*, *146*. https://doi.org/10.1016/j.metabol.2023.155640

Jeon, J., Leimbigler, P. J., Baruah, G., Li, M. H., Fossat, Y., & Whitehead, A. J. (2019). Predicting Glycaemia in Type 1 Diabetes Patients: Experiments in Feature Engineering and Data Imputation. *Journal of Healthcare Informatics Research*, *4*(1), 71–90. https://doi.org/10.1007/s41666-019-00063-2

Jeukendrup, A. (2014). A step towards personalized sports nutrition: Carbohydrate intake during exercise. *Sports Medicine (Auckland, N.Z.)*, *44 Suppl 1*(Suppl 1), S25-33. https://doi.org/10.1007/s40279-014-0148-z

Jordan, L. (2015). The problem with Big Data in Translational Medicine. A review of where we've been and the possibilities ahead. *Applied & Translational Genomics*, *6*, 3-6-3–6.

Jovanov, E. (2015). Preliminary analysis of the use of smartwatches for longitudinal health monitoring. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 865–868. https://doi.org/10.1109/EMBC.2015.7318499

Jumpertz Von Schwartzenberg, R., Vazquez Arreola, E., Sandforth, A., Hanson, R. L., & Birkenfeld, A. L. (2024). Role of weight loss-induced prediabetes remission in the prevention of type 2 diabetes: Time to improve diabetes prevention. *Diabetologia*, *67*(8), 1714–1718. https://doi.org/10.1007/s00125-024-06178-5

Kalamaras, I., Glykos, K., Megalooikonomou, V., Votis, K., & Tzovaras, D. (2021). Graph-based visualization of sensitive medical data. *Multimedia Tools and Applications*, 1-28-1–28.

Kanwal, T., Anjum, A., & Khan, A. (2021). Privacy preservation in e-health cloud: Taxonomy, privacy requirements, feasibility analysis, and opportunities. *Cluster Computing*, *24*(1), 293–317.

Karim, R. A. H., Vassányi, I., & Kósa, I. (2020). After-meal blood glucose level prediction using an absorption model for neural network training. *Computers in Biology and Medicine*, *125*, 103956. https://doi.org/10.1016/j.compbiomed.2020.103956

Kaur, K., & Rani, R. (2015). Managing data in healthcare information systems: Many models, one solution. *Computer*, *48*(3), 52-59-52–59.

Kaushik, S., Choudhury, A., Sheron, P. K., Dasgupta, N., Natarajan, S., Pickett, L. A., & Dutt, V. (2020). AI in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in Big Data*, *3*, 4–4.

Kazemi, N., Abdolrazzaghi, M., Light, P. E., & Musilek, P. (2023). In-human testing of a non-invasive continuous low-energy microwave glucose sensor with advanced machine learning capabilities. *Biosensors & Bioelectronics*, *241*, 115668. https://doi.org/10.1016/j.bios.2023.115668

Keshet, A., Reicher, L., Bar, N., & Segal, E. (2023). Wearable and digital devices to monitor and treat metabolic diseases. *Nature Metabolism*, *5*(4), 563–571. https://doi.org/10.1038/s42255-023-00778-y

Khan, M. I., Acharya, B., & Chaurasiya, R. K. (2022). Automatic Prediction of Glycemic Index Category from Food Images Using Machine Learning Approaches. *Arabian Journal for Science and Engineering*, *47*(8), 10823–10846. https://doi.org/10.1007/s13369-022-06754-0

Kim, B. S., & Yoo, S. K. (2006). Motion artifact reduction in photoplethysmography using independent component analysis. *IEEE Transactions on Bio-Medical Engineering*, *53*(3), 566–568. https://doi.org/10.1109/TBME.2005.869784

Kim, J.-C., & Chung, K. (2019). Associative feature information extraction using text mining from health big data. *Wireless Personal Communications*, *105*(2), 691–707.

Kim, M. M., Kreider, K. E., Padilla, B. I., & Lambes, K. (2022). Implementation of a Prediabetes Risk Test for an Underserved Population in a Federally Qualified Health Center. *Clinical Diabetes*, *41*(1), 102–109. https://doi.org/10.2337/cd21-0057

Kim, S., Lee, H., & Chung, Y. D. (2017). Privacy-preserving data cube for electronic medical records: An experimental evaluation. *International Journal of Medical Informatics*, *97*, 33-42-33–42.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, *26*(9), 1011–1013.

Klompas, M., Kulldorff, M., Vilk, Y., Bialek, S. R., & Harpaz, R. (2011). *Herpes zoster and postherpetic neuralgia surveillance using structured electronic data*. *86*, 1146-1153-1146–1153.

Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., Nathan, D. M., & Diabetes Prevention Program Research Group. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England Journal of Medicine*, *346*(6), 393–403. https://doi.org/10.1056/NEJMoa012512

Knutson, K. L., Spiegel, K., Penev, P., & Van Cauter, E. (2007). The metabolic consequences of sleep deprivation. *Sleep Medicine Reviews*, *11*(3), 163–178. https://doi.org/10.1016/j.smrv.2007.01.002

Krishna, S., Boren, S. A., & Balas, E. A. (2009). Healthcare via cell phones: A systematic review. *Telemedicine and E-Health*, *15*(3), 231-240-231–240.

Kroeger, J., Siegmund, T., Schubert-Olesen, O., Keuthage, W., Lettmann, M., Richert, K., & Pfeiffer, A. F. (2021). AGP and Nutrition–Analysing postprandial glucose courses with CGM. *Diabetes Research and Clinical Practice*, *174*, 108738.

Kulmanov, M., Khan, M. A., & Hoehndorf, R. (2018). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, *34*(4), 660-668-660–668.

Kumar, M., & Mishra, S. K. (2020). A comprehensive review on nature inspired neural network based adaptive filter for eliminating noise in medical images. *Current Medical Imaging*, *16*(4), 278-287-278–287.

Kumari, R., Anand, P. K., & Shin, J. (2023). Improving the Accuracy of Continuous Blood Glucose Measurement Using Personalized Calibration and Machine Learning. *Diagnostics*, *13*(15), 2514. https://doi.org/10.3390/diagnostics13152514

Kurnik Mesarič, K., Pajek, J., Logar Zakrajšek, B., Bogataj, Š., & Kodrič, J. (2023). Cognitive behavioral therapy for lifestyle changes in patients with obesity and type 2 diabetes: A systematic review and meta-analysis. *Scientific Reports*, *13*(1), 12793.

Laguna Sanz, A. J., Díez, J. L., Giménez, M., & Bondia, J. (2019). Enhanced Accuracy of Continuous Glucose Monitoring during Exercise through Physical Activity

Tracking Integration. *Sensors*, *19*(17), Article 17. https://doi.org/10.3390/s19173757

Laleci, G. B., & Dogac, A. (2009). A semantically enriched clinical guideline model enabling deployment in heterogeneous healthcare environments. *IEEE Transactions on Information Technology in Biomedicine*, *13*(2), 263-273-263–273.

Lam, B., Catt, M., Cassidy, S., Bacardit, J., Darke, P., Butterfield, S., Alshabrawy, O., Trenell, M., & Missier, P. (2021a). Using Wearable Activity Trackers to Predict Type 2 Diabetes: Machine Learning–Based Cross-sectional Study of the UK Biobank Accelerometer Cohort. *JMIR Diabetes*, *6*(1), e23364. https://doi.org/10.2196/23364

Lam, B., Catt, M., Cassidy, S., Bacardit, J., Darke, P., Butterfield, S., Alshabrawy, O., Trenell, M., & Missier, P. (2021b). Using wearable activity trackers to predict type 2 diabetes: Machine learning–based cross-sectional study of the UK Biobank accelerometer cohort. *JMIR Diabetes*, *6*(1), e23364.

Langer, P., Altmüller, S., Fleisch, E., & Barata, F. (2024). CLAID: Closing the Loop on AI & Data Collection — A cross-platform transparent computing middleware framework for smart edge-cloud and digital biomarker applications. *Future Generation Computer Systems*, *159*, 505–521. https://doi.org/10.1016/j.future.2024.05.026

Länsitie, M., Kangas, M., Jokelainen, J., Venojärvi, M., Vaaramo, E., Härkönen, P., Keinänen-Kiukaanniemi, S., & Korpelainen, R. (2021). Association between accelerometer-measured physical activity, glucose metabolism, and waist circumference in older adults. *Diabetes Research and Clinical Practice*, *178*, 108937. https://doi.org/10.1016/j.diabres.2021.108937

Lantz, E. (2016). *Machine Learning for Risk Prediction and Privacy in Electronic Health Records*. The University of Wisconsin-Madison.

Lee, S.-M., Kim, D.-Y., & Woo, J. (2023). Glucose Transformer: Forecasting Glucose Level and Events of Hyperglycemia and Hypoglycemia. *IEEE Journal of Biomedical and Health Informatics*, *27*(3), 1600–1611.

Lehmann, V., Föll, S., Maritsch, M., van Weenen, E., Kraus, M., Lagger, S., Odermatt, K., Albrecht, C., Fleisch, E., Zueger, T., Wortmann, F., & Stettler, C. (2023). Noninvasive Hypoglycemia Detection in People With Diabetes Using Smartwatch Data. *Diabetes Care*, *46*(5), 993–997. https://doi.org/10.2337/dc22-2290

Leproult, R., & Van Cauter, E. (2010). Role of sleep and sleep loss in hormonal release and metabolism. *Endocrine Development*, *17*, 11–21. https://doi.org/10.1159/000262524

Li, G., Lian, W., Qu, H., Li, Z., Zhou, Q., & Tian, J. (2021). Improving patient care through the development of a 5G-powered smart hospital. *Nature Medicine*, *27*(6), 936-937-936–937.

Li, L., Albert-Smet, I., & Faisal, A. A. (2020). Optimizing medical treatment for sepsis in intensive care: From reinforcement learning to pre-trial evaluation. *arXiv Preprint arXiv:2003.06474*.

Li, M., Fu, X., & Li, D. (2020). Diabetes prediction based on XGBoost algorithm. *IOP Conference Series: Materials Science and Engineering*, *768*(7), 072093. https://iopscience.iop.org/article/10.1088/1757-899X/768/7/072093/meta

Li, Q., Mark, R. G., & Clifford, G. D. (2008). Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. *Physiological Measurement*, *29*(1), 15–32. https://doi.org/10.1088/0967-3334/29/1/002

Liang, Y., Elgendi, M., Chen, Z., & Ward, R. (2018). An optimal filter for short photoplethysmogram signals. *Scientific Data*, *5*(1), Article 1. https://doi.org/10.1038/sdata.2018.76

Lin, F. P.-Y., Pokorny, A., Teng, C., & Epstein, R. J. (2017). TEPAPA: a novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. *Scientific Reports*, *7*(1), 1-13-1–13.

Ling, A. Y., Kurian, A. W., Caswell-Jin, J. L., Sledge Jr, G. W., Shah, N. H., & Tamang, S. R. (2019). Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. *JAMIA Open*, *2*(4), 528-537-528–537.

Little, R. R., & Sacks, D. B. (2009). HbA1c: How do we measure it and what does it mean? *Current Opinion in Endocrinology, Diabetes and Obesity*, *16*(2), 113–118.

Liu, C., Wang, F., Hu, J., & Xiong, H. (2015). *Temporal phenotyping from longitudinal electronic health records: A graph based framework*. 705-714-705–714.

Liu, J., Zhao, Y., Lai, B., Wang, H., & Tsui, K. L. (2020). Wearable Device Heart Rate and Activity Data in an Unsupervised Approach to Personalized Sleep Monitoring: Algorithm Validation. *JMIR mHealth and uHealth*, *8*(8), e18370. https://doi.org/10.2196/18370

Liu, S., Li, T., Ding, H., Tang, B., Wang, X., Chen, Q., Yan, J., & Zhou, Y. (2020). A hybrid method of recurrent neural network and graph neural network for next-period prescription prediction. *International Journal of Machine Learning and Cybernetics*, *11*(12), 2849–2856. https://doi.org/10.1007/s13042-020-01155-x

Lucas, C., Wong, P., Klein, J., Castro, T. B. R., Silva, J., Sundaram, M., Ellingson, M. K., Mao, T., Oh, J. E., Israelow, B., & others. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature*, *584*(7821), 463-469-463–469.

Lundqvist, M. H., Almby, K., Abrahamsson, N., & Eriksson, J. W. (2019). Is the Brain a Key Player in Glucose Regulation and Development of Type 2 Diabetes? *Frontiers in Physiology*, *10*. https://doi.org/10.3389/fphys.2019.00457

Maged, Y., & Atia, A. (2022). The Prediction Of Blood Glucose Level By Using The ECG Sensor of Smartwatches. *2022 2nd International Mobile, Intelligent, and*

Ubiquitous Computing Conference (MIUCC), 406–411. https://doi.org/10.1109/MIUCC55081.2022.9781730

Makarova, E., & Lagerev, D. (2020). *Methodology for Preprocessing Semi-Structured Data for Making Managerial Decisions in the Healthcare*.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, *53*(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Malik, K. M., Krishnamurthy, M., Alobaidi, M., Hussain, M., Alam, F., & Malik, G. (2020). Automated domain-specific healthcare knowledge graph curation framework: Subarachnoid hemorrhage as phenotype. *Expert Systems with Applications*, *145*, 113120–113120.

Marble, H. D., Huang, R., Dudgeon, S. N., Lowe, A., Herrmann, M. D., Blakely, S., Leavitt, M. O., Isaacs, M., Hanna, M. G., Sharma, A., & others. (2020). A regulatory science initiative to harmonize and standardize digital pathology and machine learning processes to speed up clinical innovation to patients. *Journal of Pathology Informatics*, *11*.

Maritsch, M., Föll, S., Lehmann, V., Bérubé, C., Kraus, M., Feuerriegel, S., Kowatsch, T., Züger, T., Stettler, C., Fleisch, E., & Wortmann, F. (2020). Towards Wearable-based Hypoglycemia Detection and Warning in Diabetes. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3334480.3382808

Marling, C., & Bunescu, R. (2020). The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. *CEUR Workshop Proceedings*, *2675*, 71–74.

Masoumian Hosseini, M., Masoumian Hosseini, S. T., Qayumi, K., Hosseinzadeh, S., & Sajadi Tabar, S. S. (2023). Smartwatches in healthcare medicine: Assistance and monitoring; a scoping review. *BMC Medical Informatics and Decision Making*, *23*(1), 248. https://doi.org/10.1186/s12911-023-02350-w

Mbarek, S. B., Alcheikh, N., & Younis, M. I. (2022). Recent advances on MEMS based Infrared Thermopile detectors. *Microsystem Technologies*, *28*(8), 1751–1764. https://doi.org/10.1007/s00542-022-05306-8

McCall, A. L., Lieb, D. C., Gianchandani, R., MacMaster, H., Maynard, G. A., Murad, M. H., Seaquist, E., Wolfsdorf, J. I., Wright, R. F., & Wiercioch, W. (2023). Management of individuals with diabetes at high risk for hypoglycemia: An endocrine society clinical practice guideline. *The Journal of Clinical Endocrinology & Metabolism*, *108*(3), 529–562.

McCarthy, J. J. (2005). Optimal paradigms. *Linguistics Department Faculty Publication Series*, 55–55.

McKeever, S., Ye, J., Coyle, L., Bleakley, C., & Dobson, S. (2010). Activity recognition using temporal evidence theory. *Journal of Ambient Intelligence and Smart Environments*, *2*(3), 253-269-253–269.

McKinney, W., & Team, P. D. (2015). Pandas-Powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, *1625*. http://pandas.pydata.org/pandas-docs/version/0.7.3/pandas.pdf

Melville, H., Shahid, M., Gaines, A., McKenzie, B. L., Alessandrini, R., Trieu, K., Wu, J. H. Y., Rosewarne, E., & Coyle, D. H. (2023). The nutritional profile of plant-based meat analogues available for sale in Australia. *Nutrition & Dietetics*, *80*(2), 211–222. https://doi.org/10.1111/1747-0080.12793

Menegotto, A. B., Becker, C. D. L., & Cazella, S. C. (2021). Computer-aided diagnosis of hepatocellular carcinoma fusing imaging and structured health data. *Health Information Science and Systems*, *9*(1), 1-11-1–11.

Mergenthaler, P., Lindauer, U., Dienel, G. A., & Meisel, A. (2013). Sugar for the brain: The role of glucose in physiological and pathological brain function. *Trends in Neurosciences*, *36*(10), 587–597. https://doi.org/10.1016/j.tins.2013.07.001

Miao, F., Liu, Z.-D., Liu, J.-K., Wen, B., He, Q.-Y., & Li, Y. (2019). Multi-sensor fusion approach for cuff-less blood pressure measurement. *IEEE Journal of Biomedical and Health Informatics*, *24*(1), 79-91-79–91.

Migueles, J. H., Martinez-Nicolas, A., Cadenas-Sanchez, C., Esteban-Cornejo, I., Muntaner-Mas, A., Mora-Gonzalez, J., Rodriguez-Ayllon, M., Madrid, J. A., Rol, M. A., Hillman, C. H., Catena, A., & Ortega, F. B. (2021). Activity-rest circadian pattern and academic achievement, executive function, and intelligence in children with obesity. *Scandinavian Journal of Medicine and Science in Sports*, *31*(3), 653–664. Scopus. https://doi.org/10.1111/sms.13862

Migueles, J. H., Rowlands, A. V., Huber, F., Sabia, S., & Van Hees, V. T. (2019). GGIR: A Research Community–Driven Open Source R Package for Generating Physical Activity and Sleep Outcomes From Multi-Day Raw Accelerometer Data. *Journal for the Measurement of Physical Behaviour*, *2*(3), 188–196. Scopus. https://doi.org/10.1123/jmpb.2018-0063

Miled, Z. B., Haas, K., Black, C. M., Khandker, R. K., Chandrasekaran, V., Lipton, R., & Boustani, M. A. (2020). Predicting dementia with routine care EMR data. *Artificial Intelligence in Medicine*, *102*, 101771–101771.

Ming, Y., Qu, H., & Bertini, E. (2018). Rulematrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 342-352-342–352.

Mohamed, S. M., Shalaby, M. A., El-Shiekh, R. A., El-Banna, H. A., Emam, S. R., & Bakr, A. F. (2023). Metabolic syndrome: Risk factors, diagnosis, pathogenesis, and management with natural approaches. *Food Chemistry Advances*, *3*, 100335.

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554-81542–81554.

Mohan, V., Joshi, S., Mithal, A., Kesavadev, J., Unnikrishnan, A. G., Saboo, B., Kumar, P., Chawla, M., Bhograj, A., & Kovil, R. (2023). Expert Consensus Recommendations on Time in Range for Monitoring Glucose Levels in People with Diabetes: An Indian Perspective. *Diabetes Therapy*. https://doi.org/10.1007/s13300-022-01355-4

Molugulu, N., Gubbiyappa, K. S., Murthy, C. R. V., Lumae, L., & Mruthyunjaya, A. T. (2016). Evaluation of self-reported medication adherence and its associated factors among epilepsy patients in Hospital Kuala Lumpur. *Journal of Basic and Clinical Pharmacy*, *7*(4), 105–105.

Monnier, L., Colette, C., & Owens, D. R. (2008). Glycemic Variability: The Third Component of the Dysglycemia in Diabetes. Is It Important? How to Measure It? *Journal of Diabetes Science and Technology (Online)*, *2*(6), 1094–1100.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, *65*, 211-222-211–222.

Moore, W., & Frye, S. (2019). Review of HIPAA, part 1: History, protected health information, and privacy and security rules. *Journal of Nuclear Medicine Technology*, *47*(4), 269–272.

Mouri, Mi., & Badireddy, M. (2023). Hyperglycemia. In *StatPearls [Internet]*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK430900/

Muhammad, G., Alshehri, F., Karray, F., El Saddik, A., Alsulaiman, M., & Falk, T. H. (2021). A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*.

Müller, R., Thews, O., Rohrbach, C., Sergl, M., & Pommerening, K. (1996). A graph-grammar approach to represent causal, temporal and other contexts in an oncological patient record. *Methods of Information in Medicine*, *35*(02), 127-141-127–141.

Mustafa, S., Norman, K., Kenealy, T., Paul, R., Murphy, R., Lawrenson, R., & Chepulis, L. (2024). Management of type 2 diabetes in New Zealand: A scoping review of interventions with measurable clinical outcomes. *Public Health*, *234*, 1–15.

Nabian, M., Yin, Y., Wormwood, J., Quigley, K. S., Barrett, L. F., & Ostadabbas, S. (2018). An Open-Source Feature Extraction Tool for the Analysis of Peripheral Physiological Data. *IEEE Journal of Translational Engineering in Health and*

Medicine, *6*, 1–11. IEEE Journal of Translational Engineering in Health and Medicine. https://doi.org/10.1109/JTEHM.2018.2878000

Nathan, V., & Jafari, R. (2017). Particle filtering and sensor fusion for robust heart rate monitoring using wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, *22*(6), 1834-1846-1834–1846.

Nicolaisen, S. K., Pedersen, L., Witte, D. R., Sørensen, H. T., & Thomsen, R. W. (2023). HbA1c-defined prediabetes and progression to type 2 diabetes in Denmark: A population-based study based on routine clinical care laboratory data. *Diabetes Research and Clinical Practice*, *203*, 110829.

Niskanen, J.-P., Tarvainen, M. P., Ranta-aho, P. O., & Karjalainen, P. A. (2004). Software for advanced HRV analysis. *Computer Methods and Programs in Biomedicine*, *76*(1), 73–81. https://doi.org/10.1016/j.cmpb.2004.03.004

Nouretdinov, I., Melluish, T., & Vovk, V. (2001). Ridge regression confidence machine. *ICML*, 385–392. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8d87824378c4cf198d9fde03d393b0dd5f4c99ad

Olde Bekkink, M., Koeneman, M., de Galan, B. E., & Bredie, S. J. (2019). Early Detection of Hypoglycemia in Type 1 Diabetes Using Heart Rate Variability Measured by a Wearable Device. *Diabetes Care*, *42*(4), 689–692. https://doi.org/10.2337/dc18-1843

OxWearables/biobankAccelerometerAnalysis. (2024). [Python]. Oxford Wearables Group. https://github.com/OxWearables/biobankAccelerometerAnalysis (Original work published 2014)

Paganelli, A. I., Mondéjar, A. G., da Silva, A. C., Silva-Calpa, G., Teixeira, M. F., Carvalho, F., Raposo, A., & Endler, M. (2022). Real-time data analysis in health monitoring systems: A comprehensive systematic literature review. *Journal of Biomedical Informatics*, *127*, 104009. https://doi.org/10.1016/j.jbi.2022.104009

Palanisamy, V., & Thirunavukarasu, R. (2019). Implications of big data analytics in developing healthcare frameworks–A review. *Journal of King Saud University-Computer and Information Sciences*, *31*(4), 415-425-415–425.

Park, J.-U., Kim, Y., Lee, Y., Urtnasan, E., & Lee, K.-J. (2022). A Prediction Algorithm for Hypoglycemia Based on Support Vector Machine Using Glucose Level and Electrocardiogram. *Journal of Medical Systems*, *46*(10), 68.

Patel, S. R., Malhotra, A., White, D. P., Gottlieb, D. J., & Hu, F. B. (2006). Association between Reduced Sleep and Weight Gain in Women. *American Journal of Epidemiology*, *164*(10), 947–954. https://doi.org/10.1093/aje/kwj280

Patell, R., Nigmatoulline, D., Bena, J., Messinger-Rapport, B., & Lansang, Mc. (2017). Hyperglycemia and hypoglycemia in patients with diabetes in skilled nursing facilities. *Endocrine Practice*, *23*(4), 458–465.

Pathak, V., Flatt, P. R., & Irwin, N. (2018). Cholecystokinin (CCK) and related adjunct peptide therapies for the treatment of obesity and type 2 diabetes. *Peptides*, *100*, 229–235.

Pereira, J., & Silveira, M. (2019). *Learning representations from healthcare time series data for unsupervised anomaly detection*. 1-7-1–7.

Pinter, B., Vassányi, I., Gaál, B., Mák, E., & Kozmann, Gy. (2011). Personalized Nutrition Counseling Expert System. In Á. Jobbágy (Ed.), *5th European Conference of the International Federation for Medical and Biological Engineering* (Vol. 37, pp. 957–960). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23508-5_248

Plekhanova, T., Rowlands, A. V., Davies, M. J., Hall, A. P., Yates, T., & Edwardson, C. L. (2023). Validation of an automated sleep detection algorithm using data from multiple accelerometer brands. *Journal of Sleep Research*, *32*(3), e13760. https://doi.org/10.1111/jsr.13760

Poolsup, N., Suksomboon, N., & Kyaw, A. M. (2013). Systematic review and meta-analysis of the effectiveness of continuous glucose monitoring (CGM) on

glucose control in diabetes. *Diabetology & Metabolic Syndrome*, *5*(1), 39. https://doi.org/10.1186/1758-5996-5-39

Qi, W., Wang, N., Su, H., & Aliverti, A. (2022). DCNN based human activity recognition framework with depth vision guiding. *Neurocomputing*, *486*, 261–271. https://doi.org/10.1016/j.neucom.2021.11.044

Quintero-Narvaez, C. E., & Monroy, R. (2024). Integrating Knowledge Graph Data with Large Language Models for Explainable Inference. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 1198–1199. https://doi.org/10.1145/3616855.3636507

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Raj, R., Luostarinen, T., Pursiainen, E., Posti, J. P., Takala, R. S. K., Bendel, S., Konttila, T., & Korja, M. (2019). Machine learning-based dynamic mortality prediction after traumatic brain injury. *Scientific Reports*, *9*(1), 1-13-1–13.

Ranstam, J., & Cook, J. A. (2018). LASSO regression. *Journal of British Surgery*, *105*(10), 1348–1348.

Rau, C.-S., Kuo, P.-J., Chien, P.-C., Huang, C.-Y., Hsieh, H.-Y., & Hsieh, C.-H. (2018). Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. *PloS One*, *13*(11), e0207192–e0207192.

Reis, V. M., Vianna, J. M., Barbosa, T. M., Garrido, N., Alves, J. V., Carneiro, A. L., Aidar, F. J., & Novaes, J. (2019). Are wearable heart rate measurements accurate to estimate aerobic energy cost during low-intensity resistance exercise? *PLOS ONE*, *14*(8), e0221284. https://doi.org/10.1371/journal.pone.0221284

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv Preprint arXiv:1606.05386*.

Richter, A. N., & Khoshgoftaar, T. M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine*, *90*, 1-14-1–14.

Roden, M., Petersen, K. F., & Shulman, G. I. (2024). Insulin Resistance in Type 2 Diabetes. In R. I. G. Holt & A. Flyvbjerg (Eds.), *Textbook of Diabetes* (1st ed., pp. 238–249). Wiley. https://doi.org/10.1002/9781119697473.ch17

Rodríguez-Rodríguez, I., Chatzigiannakis, I., Rodríguez, J.-V., Maranghi, M., Gentili, M., & Zamora-Izquierdo, M.-Á. (2019). Utility of Big Data in Predicting Short-Term Blood Glucose Levels in Type 1 Diabetes Mellitus Through Machine Learning Techniques. *Sensors*, *19*(20), Article 20. https://doi.org/10.3390/s19204482

Rodríguez-Rodríguez, I., Rodríguez, J.-V., Chatzigiannakis, I., & Zamora Izquierdo, M. Á. (2019). On the Possibility of Predicting Glycaemia 'On the Fly' with Constrained IoT Devices in Type 1 Diabetes Mellitus Patients. *Sensors (Basel, Switzerland)*, *19*(20), 4538. https://doi.org/10.3390/s19204538

Roglic, G. (2016). WHO Global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*, *1*(1), 3–8.

Rose, S. (2013). Mortality risk score prediction in an elderly population using machine learning. *American Journal of Epidemiology*, *177*(5), 443-452-443–452.

Salih, A., Galazzo, I. B., Raisi-Estabragh, Z., Petersen, S. E., Gkontra, P., Lekadir, K., Menegaz, G., & Radeva, P. (2021). *A new scheme for the assessment of the robustness of Explainable Methods Applied to Brain Age estimation*. 492-497-492–497.

Schrodt, J., Dudchenko, A., Knaup-Gregori, P., & Ganzinger, M. (2020). Graph-representation of patient data: A systematic literature review. *Journal of Medical Systems*, *44*(4), 1-7-1–7.

Sempionatto, J. R., Montiel, V. R.-V., Vargas, E., Teymourian, H., & Wang, J. (2021). Wearable and Mobile Sensors for Personalized Nutrition. *ACS Sensors*, *6*(5), 1745–1760. https://doi.org/10.1021/acssensors.1c00553

Sevil, M., Rashid, M., Hajizadeh, I., Park, M., Quinn, L., & Cinar, A. (2021). Physical Activity and Psychological Stress Detection and Assessment of Their Effects on Glucose Concentration Predictions in Diabetes Management. *IEEE Transactions on Biomedical Engineering*, *68*(7), 2251–2260. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/TBME.2020.3049109

Sevil, M., Rashid, M., Maloney, Z., Hajizadeh, I., Samadi, S., Askari, M. R., Hobbs, N., Brandt, R., Park, M., Quinn, L., & Cinar, A. (2020). Determining Physical Activity Characteristics From Wristband Data for Use in Automated Insulin Delivery Systems. *IEEE Sensors Journal*, *20*(21), 12859–12870. IEEE Sensors Journal. https://doi.org/10.1109/JSEN.2020.3000772

Shaheen, N., Bari, L., & Mannan, M. A. (2013). *Food composition table for Bangladesh*. University of Dhaka. http://reposit.library.du.ac.bd:8080/xmlui/handle/123456789/460

Shao, Y., Zeng, Q. T., Chen, K. K., Shutes-David, A., Thielke, S. M., & Tsuang, D. W. (2019). Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Medical Informatics and Decision Making*, *19*(1), 1-11-1–11.

Sharma, N., & Bhatt, R. (2018). Privacy preservation in WSN for healthcare application. *Procedia Computer Science*, *132*, 1243-1252-1243–1252.

Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A., & Sakharkar, M. K. (2016). Significance of HbA1c test in diagnosis and prognosis of diabetic patients. *Biomarker Insights*, *11*, BMI-S38440.

Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., & Rashidi, P. (2019). DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific Reports*, *9*(1), 1-12-1–12.

Silva, P. E., Maldaner, V., Vieira, L., de Carvalho, K. L., Gomes, H., Melo, P., Babault, N., Cipriano Jr, G., & Durigan, J. L. Q. (2018). Neuromuscular electrophysiological disorders and muscle atrophy in mechanically-ventilated

traumatic brain injury patients: New insights from a prospective observational study. *Journal of Critical Care*, *44*, 87-94-87–94.

Simpson, S., Kaufmann, M. C., Glozman, V., & Chakrabarti, A. (2020). Disease X: accelerating the development of medical countermeasures for the next pandemic. *The Lancet Infectious Diseases*, *20*(5), e108-e115-e108–e115.

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, *6*(6), 52.

Site, A., Nurmi, J., & Lohan, E. S. (2023). Machine-Learning-Based Diabetes Prediction Using Multisensor Data. *IEEE Sensors Journal*, *23*(22), 28370–28377. IEEE Sensors Journal. https://doi.org/10.1109/JSEN.2023.3319360

Song, B., Feng, Y., Li, X., Sun, Z., & Yang, Y. (2017). *Un-apriori: A novel association rule mining algorithm for unstructured EMRs*. 1-6-1–6.

Sood, S. K., & Mahajan, I. (2017). A fog-based healthcare framework for chikungunya. *IEEE Internet of Things Journal*, *5*(2), 794-801-794–801.

Spiegel, K., Leproult, R., & Van Cauter, E. (1999). Impact of sleep debt on metabolic and endocrine function. *Lancet (London, England)*, *354*(9188), 1435–1439. https://doi.org/10.1016/S0140-6736(99)01376-8

St-Onge, M.-P., Grandner, M. A., Brown, D., Conroy, M. B., Jean-Louis, G., Coons, M., Bhatt, D. L., & American Heart Association Obesity, Behavior Change, Diabetes, and Nutrition Committees of the Council on Lifestyle and Cardiometabolic Health; Council on Cardiovascular Disease in the Young; Council on Clinical Cardiology; and Stroke Council. (2016). Sleep Duration and Quality: Impact on Lifestyle Behaviors and Cardiometabolic Health: A Scientific Statement From the American Heart Association. *Circulation*, *134*(18), e367–e386. https://doi.org/10.1161/CIR.0000000000000444

Sundararajan, K., & Hees, V. van. (2020). *Sleep classification from wrist-worn accelerometer data using Random Forests*. https://doi.org/10.5281/zenodo.3752645

Sung, S.-F., Lin, C.-Y., & Hu, Y.-H. (2020). EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2922-2931-2922–2931.

Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012a). Prediabetes: A high-risk state for developing diabetes. *Lancet*, *379*(9833), 2279–2290. https://doi.org/10.1016/S0140-6736(12)60283-9

Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012b). Prediabetes: A high-risk state for diabetes development. *Lancet (London, England)*, *379*(9833), 2279–2290. https://doi.org/10.1016/S0140-6736(12)60283-9

Taghanaki, S. A., Abhishek, K., Cohen, J. P., Cohen-Adad, J., & Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: A review. *Artificial Intelligence Review*, *54*(1), 137-178-137–178.

Takkouche, B., & Norman, G. (2011). PRISMA statement. *Epidemiology*, *22*(1), 128.

Tasali, E., Leproult, R., Ehrmann, D. A., & Van Cauter, E. (2008). Slow-wave sleep and the risk of type 2 diabetes in humans. *Proceedings of the National Academy of Sciences*, *105*(3), 1044–1049. https://doi.org/10.1073/pnas.0706446105

Temko, A. (2017). Accurate Heart Rate Monitoring During Physical Exercises Using PPG. *IEEE Transactions on Biomedical Engineering*, *64*(9), 2016–2024. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/TBME.2017.2676243

Therneau, T. M., & Grambsch, P. M. (2000). The cox model. In *Modeling survival data: Extending the Cox model* (pp. 39-77-39–77). Springer.

Tran, B. X., Nghiem, S., Sahin, O., Vu, T. M., Ha, G. H., Vu, G. T., Pham, H. Q., Do, H. T., Latkin, C. A., Tam, W., & others. (2019). Modeling research topics for artificial intelligence applications in medicine: Latent Dirichlet allocation application study. *Journal of Medical Internet Research*, *21*(11), e15511–e15511.

Tuomilehto, J., Lindström, J., Eriksson, J. G., Valle, T. T., Hämäläinen, H., Ilanne-Parikka, P., Keinänen-Kiukaanniemi, S., Laakso, M., Louheranta, A., Rastas, M., Salminen, V., Uusitupa, M., & Finnish Diabetes Prevention Study Group. (2001). Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *The New England Journal of Medicine*, *344*(18), 1343–1350. https://doi.org/10.1056/NEJM200105033441801

Turner-McGrievy, G. M., Dunn, C. G., Wilcox, S., Boutté, A. K., Hutto, B., Hoover, A., & Muth, E. (2019). Defining Adherence to Mobile Dietary Self-Monitoring and Assessing Tracking Over Time: Tracking at Least Two Eating Occasions per Day Is Best Marker of Adherence within Two Different Mobile Health Randomized Weight Loss Interventions. *Journal of the Academy of Nutrition and Dietetics*, *119*(9), 1516–1524. https://doi.org/10.1016/j.jand.2019.03.012

Twomey, N., Diethe, T., Fafoutis, X., Elsts, A., McConville, R., Flach, P., & Craddock, I. (2018). A Comprehensive Study of Activity Recognition Using Accelerometers. *Informatics*, *5*(2), Article 2. https://doi.org/10.3390/informatics5020027

Umpierrez, G. E., Davis, G. M., ElSayed, N. A., Fadini, G. P., Galindo, R. J., Hirsch, I. B., Klonoff, D. C., McCoy, R. G., Misra, S., Gabbay, R. A., Bannuru, R. R., & Dhatariya, K. K. (2024). Hyperglycemic Crises in Adults With Diabetes: A Consensus Report. *Diabetes Care*, *47*(8), 1257–1275. https://doi.org/10.2337/dci24-0032

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, *64*(7), 456-464-456–464.

van den Brink, W., Bloem, R., Ananth, A., Kanagasabapathi, T., Amelink, A., Bouwman, J., Gelinck, G., van Veen, S., Boorsma, A., & Wopereis, S. (2021). Digital Resilience Biomarkers for Personalized Health Maintenance and Disease

Prevention. *Frontiers in Digital Health*, *2*. https://doi.org/10.3389/fdgth.2020.614670

van Doorn, W. P. T. M., Foreman, Y. D., Schaper, N. C., Savelberg, H. H. C. M., Koster, A., van der Kallen, C. J. H., Wesselius, A., Schram, M. T., Henry, R. M. A., Dagnelie, P. C., de Galan, B. E., Bekers, O., Stehouwer, C. D. A., Meex, S. J. R., & Brouwers, M. C. G. J. (2021). Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study. *PLoS ONE*, *16*(6), e0253125. https://doi.org/10.1371/journal.pone.0253125

van Doorn, W. P. T. M., Stassen, P. M., Borggreve, H. F., Schalkwijk, M. J., Stoffers, J., Bekers, O., & Meex, S. J. R. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS One*, *16*(1), e0245157–e0245157.

van Gent, P., Farah, H., van Nes, N., & van Arem, B. (2019). HeartPy: A novel heart rate algorithm for the analysis of noisy signals. *Transportation Research Part F: Traffic Psychology and Behaviour*, *66*, 368–378. https://doi.org/10.1016/j.trf.2019.09.015

van Hees, V. T., Sabia, S., Jones, S. E., Wood, A. R., Anderson, K. N., Kivimäki, M., Frayling, T. M., Pack, A. I., Bucan, M., Trenell, M. I., Mazzotti, D. R., Gehrman, P. R., Singh-Manoux, B. A., & Weedon, M. N. (2018). Estimating sleep parameters using an accelerometer without sleep diary. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-31266-z

Venkatesh, R., Balasubramanian, C., & Kaliappan, M. (2019). Development of big data predictive analytics model for disease prediction using machine learning technique. *Journal of Medical Systems*, *43*(8), 1-8-1–8.

Vest, J. R., Grannis, S. J., Haut, D. P., Halverson, P. K., & Menachemi, N. (2017). Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *International Journal of Medical Informatics*, *107*, 101-106-101–106.

Vettoretti, M., Cappon, G., Facchinetti, A., & Sparacino, G. (2020). Advanced Diabetes Management Using Artificial Intelligence and Continuous Glucose Monitoring Sensors. *Sensors (Basel, Switzerland)*, *20*(14), 3870. https://doi.org/10.3390/s20143870

Visani, G., Bagli, E., & Chesani, F. (2020). OptiLIME: Optimized LIME explanations for diagnostic computer algorithms. *arXiv Preprint arXiv:2006.05714*.

Voss, A., Schroeder, R., Heitmann, A., Peters, A., & Perz, S. (2015). Short-Term Heart Rate Variability—Influence of Gender and Age in Healthy Subjects. *PLOS ONE*, *10*(3), e0118308. https://doi.org/10.1371/journal.pone.0118308

Walch, O. (2019). Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography. *PhysioNet*, *101*.

Walch, O., Huang, Y., Forger, D., & Goldstein, C. (2019). Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, *42*(12). Scopus. https://doi.org/10.1093/sleep/zsz180

Wallace, B. C., Kuiper, J., Sharma, A., Zhu, M., & Marshall, I. J. (2016). Extracting PICO sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, *17*(1), 4572–4596.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*(3), 230-238-230–238.

Wang, F., Cui, P., Pei, J., Song, Y., & Zang, C. (2020). *Recent Advances on Graph Analytics and Its Applications in Healthcare*. 3545-3546-3545–3546.

Wang, J., Ji, J., Zhang, M., Lin, J.-W., Zhang, G., Gong, W., Cen, L.-P., Lu, Y., Huang, X., Huang, D., & others. (2021). Automated Explainable Multidimensional Deep Learning Platform of Retinal Images for Retinopathy of Prematurity Screening. *JAMA Network Open*, *4*(5), e218758-e218758-e218758–e218758.

Wang, L., Fan, R., Zhang, C., Hong, L., Zhang, T., Chen, Y., Liu, K., Wang, Z., & Zhong, J. (2020). Applying Machine Learning Models to Predict Medication Nonadherence in Crohn's Disease Maintenance Therapy. *Patient Preference and Adherence*, *14*, 917–917.

Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., Wang, Y., Douiri, A., Wolfe, C. D., & Bray, B. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PloS One*, *15*(6), e0234722–e0234722.

Wang, Y.-K., & Chen, C.-Y. (2023). Integrating Mobile Devices and Wearable Technology for Optimal Sleep Conditions. *Applied Sciences (Switzerland)*, *13*(17). Scopus. https://doi.org/10.3390/app13179921

Weitschies, W., Müller, L., Grimm, M., & Koziolek, M. (2021). Ingestible devices for studying the gastrointestinal physiology and their application in oral biopharmaceutics. *Advanced Drug Delivery Reviews*, 113853–113853.

Whitney, V. K. M. (1974). *Relational data management implementation techniques*. 321–350.

Wibawa, F., Catak, F. O., Kuzlu, M., Sarp, S., & Cali, U. (2022). *Homomorphic Encryption and Federated Learning based Privacy-Preserving CNN Training: COVID-19 Detection Use-Case*. 85–90.

Winter, A., Brigl, B., & Wendt, T. (2003). Modeling hospital information systems (part 1): The revised three-layer graph-based meta model 3LGM2. *Methods of Information in Medicine*, *42*(05), 544-551-544–551.

Wolkowicz, K. L., Aiello, E. M., Vargas, E., Teymourian, H., Tehrani, F., Wang, J., Pinsker, J. E., Doyle, F. J., Patti, M., Laffel, L. M., & Dassau, E. (2020). A review of biomarkers in the context of type 1 diabetes: Biological sensing for enhanced glucose control. *Bioengineering & Translational Medicine*, *6*(2), e10201. https://doi.org/10.1002/btm2.10201

Wright, A., McEvoy, D. S., Aaron, S., McCoy, A. B., Amato, M. G., Kim, H., Ai, A., Cimino, J. J., Desai, B. R., El-Kareh, R., & others. (2019). Structured override

reasons for drug-drug interaction alerts in electronic health records. *Journal of the American Medical Informatics Association*, *26*(10), 934-942-934–942.

Wu, M., Cao, H., Nguyen, H.-L., Surmacz, K., & Hargrove, C. (2015). Modeling perceived stress via HRV and accelerometer sensor streams. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1625–1628. https://doi.org/10.1109/EMBC.2015.7318686

Wu, T., Wang, Y., Wang, Y., Zhao, E., & Yuan, Y. (2021). Leveraging graph-based hierarchical medical entity embedding for healthcare applications. *Scientific Reports*, *11*(1), 1-13-1–13.

Xiao, Q., Moore, S. C., Keadle, S. K., Xiang, Y.-B., Zheng, W., Peters, T. M., Leitzmann, M. F., Ji, B.-T., Sampson, J. N., Shu, X.-O., & Matthews, C. E. (2016). Objectively measured physical activity and plasma metabolomics in the Shanghai Physical Activity Study. *International Journal of Epidemiology*, *45*(5), 1433–1444. Scopus. https://doi.org/10.1093/ije/dyw033

Xie, F., Yuan, H., Ning, Y., Ong, M. E. H., Feng, M., Hsu, W., Chakraborty, B., & Liu, N. (2022). Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of Biomedical Informatics*, *126*, 103980. https://doi.org/10.1016/j.jbi.2021.103980

Xu, R., Baracaldo, N., Zhou, Y., Anwar, A., & Ludwig, H. (2019). *Hybridalpha: An efficient approach for privacy-preserving federated learning*. 13–23.

Xu, Y., Hübener, I., Seipp, A.-K., Ohly, S., & David, K. (2017). From the lab to the real-world: An investigation on the influence of human movement on Emotion Recognition using physiological signals. *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 345–350. https://doi.org/10.1109/PERCOMW.2017.7917586

Xu, Z., Chou, J., Zhang, X. S., Luo, Y., Isakova, T., Adekkanattu, P., Ancker, J. S., Jiang, G., Kiefer, R. C., & Pacheco, J. A. (2020). Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks. *Journal of Biomedical Informatics*, *102*, 103361.

Yeung, R. O., Retnakaran, R., Savu, A., Butalia, S., & Kaul, P. (2024). Gestational diabetes: One size does not fit all—an observational study of maternal and neonatal outcomes by maternal glucose profile. *Diabetic Medicine*, *41*(2), e15205. https://doi.org/10.1111/dme.15205

Yoo, H., & Chung, K. (2018). Heart rate variability based stress index service model using bio-sensor. *Cluster Computing*, *21*(1), 1139–1149. https://doi.org/10.1007/s10586-017-0879-3

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., Lin, S. M., Zhang, W., Zhang, P., & Sun, H. (2020). Graph embedding on biomedical networks: Methods, applications and evaluations. *Bioinformatics*, *36*(4), 1241-1251-1241–1251.

Yun, J.-S., & Ko, S.-H. (2015). Avoiding or coping with severe hypoglycemia in patients with type 2 diabetes. *The Korean Journal of Internal Medicine*, *30*(1), 6.

Zahedani, A. D., McLaughlin, T., Veluvali, A., Aghaeepour, N., Hosseinian, A., Agarwal, S., Ruan, J., Tripathi, S., Woodward, M., Hashemi, N., & Snyder, M. (2023). Digital health application integrating wearable data and behavioral patterns improves metabolic health. *Npj Digital Medicine*, *6*(1), Article 1. https://doi.org/10.1038/s41746-023-00956-y

Zhang, J. (2007). Effect of Age and Sex on Heart Rate Variability in Healthy Subjects. *Journal of Manipulative and Physiological Therapeutics*, *30*(5), 374–379. https://doi.org/10.1016/j.jmpt.2007.04.001

Zhang, Q., Zhou, J., & Zhang, B. (2020). Graph based multichannel feature fusion for wrist pulse diagnosis. *IEEE Journal of Biomedical and Health Informatics*.

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of Translational Medicine*, *4*(2), 30. https://doi.org/10.3978/j.issn.2305-5839.2015.12.63

Zhao, J., Lv, Y., Zeng, Q., & Wan, L. (2024). Online Policy Learning-Based Output-Feedback Optimal Control of Continuous-Time Systems. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *71*(2), 652–656. IEEE Transactions on

Circuits and Systems II: Express Briefs. https://doi.org/10.1109/TCSII.2022.3211832

Zhao, M., Havrilla, J., Peng, J., Drye, M., Fecher, M., Guthrie, W., Tunc, B., Schultz, R., Wang, K., & Zhou, Y. (2022). Development of a phenotype ontology for autism spectrum disorder by natural language processing on electronic health records. *Journal of Neurodevelopmental Disorders*, *14*(1), 1–12.

Zhao, X., An, X., Yang, C., Sun, W., Ji, H., & Lian, F. (2023). The crucial role and mechanism of insulin resistance in metabolic disease. *Frontiers in Endocrinology*, *14*, 1149239.

Zhu, T., Uduku, C., Li, K., Herrero, P., Oliver, N., & Georgiou, P. (2022). Enhancing self-management in type 1 diabetes with wearables and deep learning. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00626-5

Zimmet, P., Alberti, K. G., Magliano, D. J., & Bennett, P. H. (2016). Diabetes mellitus statistics on prevalence and mortality: Facts and fallacies. *Nature Reviews Endocrinology*, *12*(10), 616–622.

Zoungas, S., Chalmers, J., Ninomiya, T., Li, Q., Cooper, M. E., Colagiuri, S., Fulcher, G., de Galan, B. E., Harrap, S., Hamet, P., Heller, S., MacMahon, S., Marre, M., Poulter, N., Travert, F., Patel, A., Neal, B., Woodward, M., & ADVANCE Collaborative Group. (2012). Association of HbA1c levels with vascular complications and death in patients with type 2 diabetes: Evidence of glycaemic thresholds. *Diabetologia*, *55*(3), 636–643. https://doi.org/10.1007/s00125-011-2404-1

Zuallaert, J., Godin, F., Kim, M., Soete, A., Saeys, Y., & De Neve, W. (2018). SpliceRover: Interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics*, *34*(24), 4180-4188-4180–4188.

Zuo, Z., Li, J., Xu, H., & Al Moubayed, N. (2021). Curvature-based feature selection with application in classifying electronic health records. *Technological Forecasting and Social Change*, *173*, 121127.

Appendix

Paper 1

electronics

MDPI

Review

Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning

Haider Ali [1], Imran Khan Niazi [1,2,3], Brian K. Russell [1,4], Catherine Crofts [1], Samaneh Madanian [1] and David White [1,*]

[1] BioDesign Lab, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand
[2] Centre for Chiropractic Research, New Zealand College of Chiropractic, Auckland 1060, New Zealand
[3] Center for Sensory-Motor Interaction, Department of Health Science and Technology, Aalborg University, 9220 Aalborg Øst, Denmark
[4] Ambient Cognition Limited, Auckland 1010, New Zealand
* Correspondence: david.white@biodesignlab.co.nz; Tel.: +64-211-23-5470

Abstract: Electronic medical records (EMRs) help in identifying disease archetypes and progression. A very important part of EMRs is the presence of time domain data because these help with identifying trends and monitoring changes through time. Most time-series data come from wearable devices monitoring real-time health trends. This review focuses on the time-series data needed to construct complete EMRs by identifying paradigms that fall within the scope of the application of artificial intelligence (AI) based on the principles of translational medicine. (1) Background: The question addressed in this study is: What are the taxonomies present in the field of the application of machine learning on EMRs? (2) Methods: Scopus, Web of Science, and PubMed were searched for relevant records. The records were then filtered based on a PRISMA review process. The taxonomies were then identified after reviewing the selected documents; (3) Results: A total of five main topics were identified, and the subheadings are discussed in this review; (4) Conclusions: Each aspect of the medical data pipeline needs constant collaboration and update for the proposed solutions to be useful and adaptable in real-world scenarios.

Keywords: time series; electronic medical records; systemic review; artificial intelligence; machine learning

Citation: Ali, H.; Niazi, I.K.; Russell, B.K.; Crofts, C.; Madanian, S.; White, D. Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning. *Electronics* **2023**, *12*, 554. https://doi.org/10.3390/electronics12030554

Academic Editors: Mohammed Abdulhakim Al-Absi and Rui Fu

Received: 13 December 2022
Revised: 15 January 2023
Accepted: 19 January 2023
Published: 21 January 2023

1. Introduction

Translational medicine (TM) includes collaboration between medical clinicians, biomedical engineers, and scientists to develop artificial intelligence (AI) models that account for differing data sources, data collection methods, and other real-world factors. In some cases, TM reduces the time from solution development to deployment. It does so by allowing effective communication between different players for shared goals. It is characterized by integrating digital biomarkers, multi-omics profiling, model-based data, AI, biomarker-guided trial designs, and patient-centric companion diagnostics. Therefore, the taxonomies identified in this review are guided by translational medicine [1]. Ref [2] presents the following components of the complete medical record including: connected fitness devices, patient-focused personal health records, individual behavioral patterns, pharmacy-focused medical adherence data, provider-focused medical records, connected medical devices, and genomic information.

In these complete electronic medical records (EMRs), the use of time-series data is essential because most of the biomarkers are tracked as trends in time [3]. EMRs help in disease monitoring, pandemic monitoring, adjustment of lifestyles, hospitals, intensive care units, and integration of healthcare services. Machine learning (ML) is "The ability of computers to advise decisions based on the available data" [4]. ML has been used in

applied clinical studies for some time now [5], and recent renewed interest has been driven by increased data availability [6] and an increase in computational capacity [7]. The most common medical data types are images, such as computerized tomography (CT) scans, time-series data, and blood, urine, and metabolic panels. Some noteworthy literature reviews cover computer vision techniques to assist clinical decision-making [8–10]. However, there is a need to review the broader research landscape relating to the application of ML in time-series data in EMRs to give the relevant taxonomies, patterns in literature, emerging trends, and knowledge streams in this field. Time series can be defined as repeated readings from a device over a period of time. The frequency of the readings can be periodic or non-periodic and range from thousands of times per second (e.g., accellerometry) to a few times a day (e.g., glucose monitoring).

Our review of the relevant studies has found numerous pertinent publications [11], including a previous work by Davy et al. (2015) that investigates the effectiveness of chronic care AI models employed in primary healthcare [12]. A more recent work by Chen et al. (2020) reviews the use of probabilistic ML models applied to healthcare data [13], while Wang et al. (2021) examined the latest advancements in graph-based analytics in healthcare [14]. The application of telemedicine in maintaining electronic health records was recently reviewed by Gu et al. (2019) using a cite space analysis [15]. Cite space was also used in a scientometric review of the application of latent discriminant analysis in healthcare data by Tean et al. (2019) [16]. Emerging challenges when using unstructured EMRs were deliberated by Adnan et al. (2020) in the context of using big unstructured data in healthcare [17].

In comparison to the studies above our study has three novel features:

1. It identifies taxonomies within the field after a systemic search of research databases.
2. It finds these taxonomies based on the principles of translational medicine so that the reader may find all the information needed for a translational solution in one place.
3. It identifies the core challenges and advancements in each taxonomy and provides a rigorous volume of the literature to serve as a baseline.

The following are the research questions we will try to address:

1. What paradigms fall under the umbrella of AI in time series and graph-based healthcare data?
2. What are the latest advances in these domains?
3. What are the latest challenges in these taxonomies?

Although earlier literature reviews have covered the application of specific algorithms and problems in EMRs, as shown in Table 1, a need exists to develop taxonomies and discuss recent findings.

Table 1. Comparison with earlier works.

References	Time Series	Disease Specific	Translational Medicine
[12]	✓	X	X
[18]	✓	✓	X
[19]	X	✓	X
[16]	X	✓	X
[15]	✓	✓	X
[17]	✓	✓	X
Our work	✓	X	✓

2. Materials and Methods

A paradigm is a collection of elements based on a common lexeme [20]. Identifying the paradigms of fields interacting with each other is imperative to achieving overarching solutions. In this review, these paradigms are based on an industry perspective of TM [1] and TM principles. The three Ts in TM are developing treatments and interventions, testing

the proposed interventions' effectiveness, and deploying these applications in the real world [21]. In this review, all the paradigms in the field considered are presented by Figure 1, and the challenges in applying AI-based solutions to TM are also discussed.

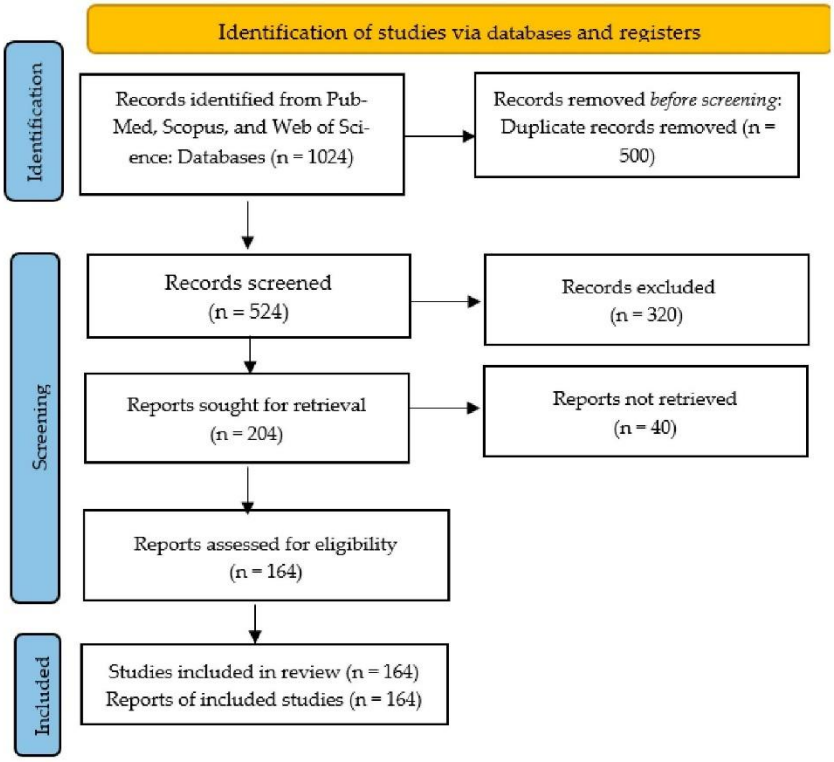


Figure 1. PRISMA review process for selection of records from research databases.

In this review, the literature was searched from the following databases: PubMed, Scopus, and Web of Science, and the papers between 2015–2022 were selected. The search terms were (("Physiological sensors" OR "Biomedical sensors" OR "Bio-medical Sensors") AND ("Machine Learning" OR "ML" OR "Artificial Intelligence" OR "AI" OR "Deep learning" OR "DL" OR "Reinforcement learning" OR "Electronic health records") NOT ("Security and Privacy") NOT ("images") NOT ("Robot")). The search was limited from 2015 to November 2022. Only articles were included, and reviews were not made part of the search criteria. The proceedings of various conferences were excluded from the search. Applying these filters and only selecting English language records resulted in 320 articles being removed, resulting in the number of articles selected for review totaling 164. These records were then collected and thoroughly read to answer the following questions:

- What is the type of data?
- What kind of algorithm is used?
- What pre-processing methods are used?
- What post-processing methods are used?
- What data privacy standards are observed?

- What interoperability or fusion techniques are used?

By answering these questions in the form of a table, the top three methods are identified within each paradigm. However, if the top three topics in a paradigm are repeated in any other paradigm, the next three topics are also discussed in the section to give a holistic overview of the topic. For example, the top three topics in the subtopic: time-series data and structured data are the same and, therefore, positioned in the structured data section.

An Ishikawa Fishbone Diagram (Figure 2) presents the different paradigms on the topic.

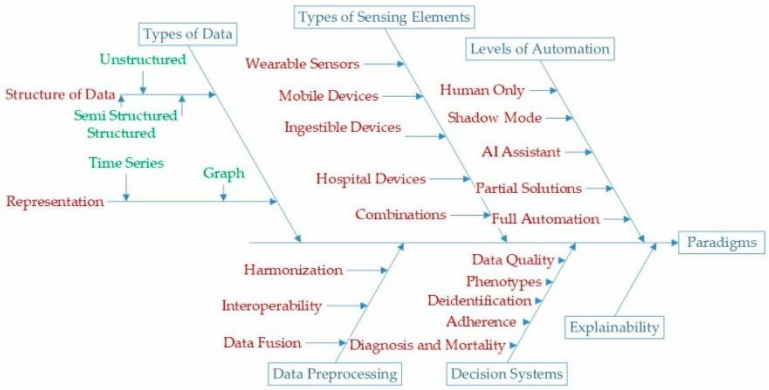


Figure 2. Ishikawa Fishbone diagram of the paradigms.

Figure 2 is a representation of the paradigms found in the literature under the topic at hand. The inclined axes represent the major topics in the field, for example, types of data and levels of automation. These inclined axes are also divided, shown as horizontal arrows, to identify the relevant subtopics.

3. Results

After collecting the records through the process previously described, then reading their methods and results, using TM principles, we enabled the paradigms to then be identified. The subsequent discussion is arranged as follows; first, we identify the most commonly occurring topics in the records that we have collected within each paradigm, and then the challenges within the paradigms are elaborated upon. This review is a combination of a narrative review built on systematic data collection.

3.1. Types of Data

One of the most obvious choices of paradigms is the type of data, which can depend upon the source of data (different types of sensors), organization of data (structured, unstructured), and representation of data (time series, images, graph representations).

Representation of Data

The way that the data are represented makes them suitable for a particular type of analysis to then be undertaken. The inductive biases of different ML models can be matched to different types of data representation. For example, as we will see in later parts of this review, the graph-based data representations are widely used for phenotyping (phenotype of the collection of observable effects of a disease), since the inductive bias of a graph-based ML model suits the relational nature of the problem of phenotyping. Thus, the first axis, along which we separated the data, is according to how it is being represented. This can take tabular, time series, or graph forms.

1. Time series (tabular representation): Time-series data contain information from physiological events in the form of time-varying biomarkers. Three leading solutions are specific to this data type: motif or pattern detection, data generation/imputation, and time-series forecasting. Generative AI models can potentially overcome the lack of access to time-series data by synthetically producing the missing and unknown data; however, accuracy by patient needs to be proven for each application, where missing data can be missing readings within a time series or the complete absence of a time-series recording in the EHR. Guided evolutionary networks (GENs) combine artificial neural networks and optimization algorithms such as genetic algorithms. These are used to fuse various information sources [22,23]. GENs are also used to discover time-series motifs in ECG data [24]. Ref. [25] uses a multilayer perceptron for time-series forecasting in healthcare data. The following Table 2 presents a comparison of the representative literature.

Table 2. Comparison of the time-series solution.

References	Applications	Sensors	Generative	Predictive	Clinical	Imputation
[26]	Motif Discovery	ECG	✓	X	✓	X
[22]	Motif Discovery	ECG and EEG	✓	✓	X	✓
[24]	Anomaly Detection	ECG	X	✓	✓	X
[25]	Expenditure Calculation	Healthcare data	X	✓	✓	X
[27]	Benchmarking	MIMIC-III	X	✓	✓	X
[28]	Imputation	ECG, MIMIC	X	✓	✓	✓

2. Graph representation: Healthcare data are relational, which makes them suitable for graphical representation. Relational data are characterized by the relations or dependence that exists amongst the rows and columns [29]. Graph-based techniques are used for developing graph-based representations of healthcare data, identifying clinical pathways and phenotypes of disease, and performing predictive modelling of disease and interventions. For example, refs. [30,31] are some typical graph representations of healthcare data. Ref. [32] determines the temporal phenotypes based on graph representations of healthcare data. Ref. [33] is a fog-based temporal network graph analysis for the Chikungunya virus in India. Ref. [34] uses a proximity-preserving graph embedding to represent electronic health records for hypertension. Ref. [35] incorporates metadata of the patients along with their vitals and lab results to learn a graph representation of electronic healthcare data. Ref. [36] is a study that employs cryptographic techniques for information embedding in the healthcare data. Ref. [37] is another knowledge-graph-based phenotyping technique for subarachnoid hemorrhage. Ref. [38] is a graph-based visualization for sensitive outcomes in medicine for healthcare data. Ref. [39] is a graph-based channel fusion for wrist pulse detection. Ref. [40] uses graphs for learning a lower dimensional representation of drug–disease interaction. As illustrated in [41], the main applications of graphs in medical interventions are drug–drug interaction, drug–disease interaction, protein–protein interaction, medical term classification, and protein function prediction. The three main methods to realize these ends are matrix factorization, random walk, and neural network-based methods. These include Laplacian methods, as demonstrated in [42], deep walk methods, as shown in [43], and neural networks, as illustrated in [40]. Graph algorithms commonly used can be categorized into temporal data mining [44], causal and contextual [45], and patient enteric graphs [46]. It is worth noting that there is no unique graph representation for sensor data or electronic medical records. Hence, most research focuses on developing graph-based presentations. One crucial research area is benchmarking and creating a numeric qualitative marker of adequate representation. There are several limitations of time-series- and graph-based healthcare

data; these include data sparsity [47], noise [48], limited generalizability [49], and lack of context [49].

The following Table 3 presents a comparison of the representative literature.

Table 3. Comparison of graph-based solutions.

References	Application	Techniques Used	Data	Contributions	Predictive	Descriptive
[32]	Temporal Phenotyping	Attention Models	MIMIC-III	10% greater than RNN in disease prediction and 3% improved areas under ROC	✓	✓
[38]		Hinge Loss		Predicted congestive health failure with an 80% accuracy. The area under the curve for patient readmission increased by over 50% from the spectral clustering	✓	✓
[36]	Graph representation	Note Binning	STRIDE	Developed term and concept mappings	X	✓
[39]	Feature fusion	Multi-Channel feature fusion	Pressure and Photo–electric Sensors	93.1% accuracy in predicting diabetes from pulse detection data.	X	X

3.2. *Structure of Data*

Another way to classify the type of healthcare data is the structure of available data. Most healthcare data are not structured against a set of rules. The structure of data dictates the kind of preprocessing required or the kind of algorithms that can be used.

A. Structured data: This follows a definite set of rules or schemes [50]. The main issues when using ML and structured data are data generation, data fusion, pattern detection, privacy preservation, and prediction of outcomes. Privacy preservation is guided by HIPAA rules [51]. Generative algorithms are used extensively to impute the missing data in the structured datasets [52,53]. Data fusion is another typical application of ML for combining two different kinds of structured data [54,55]. Federated learning that trains the models based on data from various decentralized devices is used extensively for privacy preservation of healthcare data [56–59]. ML and structured data are also valuable in predicting the outcomes of interventions, for example, [60] analyzes the user's choice in the event of alerts from clinical decision systems for potential drug–drug interference. Ref. [61] uses structured and unstructured data to find the social determinants of health characterized by social behavior, demographic features, and environmental factors of medical status and health care access. Ref. [62] is a systemic review of records from PubMed and Web of Science on the detection of strokes from structured data that found the leading keyword to be mortality and the most used algorithms to be neural networks, support vector machines, and XGBoost. Ref. [63] is another review that looked at the statistical and predictive machine learning models for cancer risk and found the cox model [64] is the most commonly used algorithm for predicting disease onset based on the input features. Ref. [65] used AI to auto-complete structured clinical records based on context. Ref. [66] is a model to detect probable cases of dementia using structured and unstructured data that uses a latent Dirichlet algorithm for feature extraction and a logistic regression model. The key issues of research for structured data in healthcare are detecting phenotypes from electronic health records [67,68], privacy and encoding of information [69–72], data harmonization from various sources [72],

synthetic data generation for research [73–75], and fairness and bias in the structured data [76].

B. Semi-structured data: These EMRs have no specific structure, enabling categorical data, meta-data, and numerical data to be entered in any field. The key areas in application of ML in unstructured data is in the conversion to structured data, predictive modeling, and interoperability of different kinds of data sources. For example, an application of ML with unstructured data for predictive modeling is used [77] to derive contextual information to generate semi-structured data from electronic medical records. Ref. [78] is a method to allocate resources from the knowledge of semi-structured healthcare data. Ref. [79] uses HL7 standards to develop the interoperability of structured, semi-structured, and unstructured data to develop obesity phenotypes. Ref. [80] is another such system that uses open EMRs to this end. Ref. [81] detects autism from semi-structured and unstructured data using a combination of skip-gram models.

C. Unstructured data: Most EMRs are unstructured [82]. Key research areas for ML applications in unstructured data are conversion amongst the various kinds of data structure and predictive modeling. An example of predictive modeling using unstructured data [83] employs unstructured EMRs to phenotype depression in youth. Latent Dirichlet Analysis (LDA) and other dimensionality reduction methods are used to obtain the hidden information between different kinds of data and then leverage it for predictive modeling [84–87]. A priori algorithms and other Bayesian methods are used to convert unstructured data to structured data [37,88,89], and in so doing, these works can also combine with structured data to make predictions [90,91]. Another technique that is relevant to the conversion of unstructured data to structured data is distant supervision. Distant supervision is a method for labeling the data by utilizing the known structures of similar data [92,93]. Exploratory text analysis is also used for pattern analysis for predictive modeling in this [94,95].

The following Table 4 compares these techniques:

Table 4. Comparison of unstructured data.

References	Application	Techniques Used	Evaluation Metrics	Structured Data
[83]	Detection of clinical depression	NLP	Specificity:97%. Sensitivity:45%	X
[84]	Disease prediction	LDA	AUC 0.94, Sensitivity 0.87 and Specificity 0.87	✓
[94]	HPV detection	NLP	AUC: 0.861	X
[92]	Breast cancer detection	NLP	AUC 0.91, Sensitivity: 0.861, Specificity 0.878, Accuracy 0.870.	✓

The different ML techniques used in conjunction with unstructured data are clustering, classification, boosting, and a combination of these three. Clustering can help with phenotyping and grouping together different clinical pathways. Classification requires labeling the data, which can be taxing for a large volume of clinical notes. Boosting models can leverage the different structures present in unstructured data to make meaningful predictions, especially, risk and mortality.

Natural language processing techniques are extensively applied to unstructured data to detect disease onset. Data harmonization and standardization is also an essential topic of discussion in unstructured healthcare.

In the healthcare context, structured and semi-structured data are typically easier to work with and analyze because they have some inherent structure. Unstructured data, such as free-text notes in electronic health records, can be more challenging to work with because they require more effort to extract meaningful information.

3.3. Types of Sensing Elements

Types of data are dependent on the types of sensing elements used. There are many types of sensing elements including wearable sensors, mobile device sensors, ingestible sensors, medical devices from hospitals, and a combination of all or some of the factors mentioned earlier.

- Wearable sensors: These bridge the gap between assessment and onset prediction. The data sources measure the biomarkers from the physiological signals in real-time, making this a vital component of multi-omics profiling [96].
- Mobile devices: Along with real-time monitoring using mobile sensors, mobile devices also allow for input from the user, making them helpful in tracking medical adherence [97].
- Ingestible sensors: Drug adherence [98] and monitoring [99] are some applications of ingestible sensors.
- Medical devices from hospitals: These include connected medical devices intended to enhance healthcare quality for people in the hospital [100].
- Combinations: The combination of the sensors enables the Internet of Medical Devices [101].

The critical limitations of wearable sensors are the contextualization of data and integration with the existing clinical care pathways; hence, a challenge exists to show clinical efficacy. Most historic clinical data are taken from a patient at rest (e.g., resting heart rate) with the assumption that only disease can shift homeostasis, and most wearable data are ambulatory (e.g., heart rate during a workout) with confounders such as physical activity making traditional clinical interpretation challenging. Interoperability is another crucial aspect that needs to be addressed when deciding on different sensing elements. This will help increase the generalizability of models by allowing them access to various kinds of data.

3.4. Data Preprocessing

As we have seen previously, data can come from various sources and in various forms. For the successful application of ML, these data must be harmonized and standardized. Data harmonization standards and intelligent interoperability techniques are the two classes along the knowledge stream. Another axis to classify data preprocessing techniques is the data fusion methods, which include feature level, data level, and decision level fusion. One more way to organize the data prior to analysis is through preprocessing techniques. These include filtering, feature extraction, and natural language processing techniques.

1. Data harmonization standards: These standards describe the preprocessing technique that prepares different kinds of data to become compatible with each other. It allows the AI to access a diversity of information through access to researcher and institution knowledge [102]. Some standards are specific to the medical cases they deal with [103–105]; however, there exists a set of medical means to ensure interoperability. The most common standards are Health Level 7 (HL7), openEHR, and ISO/IEEE 11073 Personal Health Data (PHD) standards [106], International Statistical Classification of Diseases version 10 (ICD-10) [107] and Current Procedural Terminology (CPT) codes [108]
2. Intelligent interoperability: Here, ML or other algorithms are used to combine the information from different data sources, and particularly EMRs. In intelligent interoperability of healthcare components, artificial intelligence or some other rule-based systems are used to automatically draw the relevant information from the EMRs or

sensor data. These systems use different algorithms to ensure the interoperability of various data sources. The following Table 5 elucidates such strategies. Although these systems allow for effective data communication while ensuring information integrity, one key issue is allowing for the encoding of categorical features so that the information is stored effectively.

Table 5. Comparison of interoperability techniques.

Name	Properties	References
Blockchain technology	Focused on patients rather than healthcare providers. Data are linked to the patient, aggregated, and then sensitive information such as allergies is published on the blockchain, ensuring privacy and data immutability.	[109,110]
Internet of Things	It employs the principles of the internet of things for data interoperability. It uses the protocols of Message Queuing Telemetry Transport (MQTT) to publish the relevant patient information.	[111]
Dynamic Semantic Web services	It uses the dynamic semantic web to convert the data into the HL7 framework.	[112]
Cloud Based Interoperability	It uses cloud-based models, for example, amazon web services, Microsoft Azure, and IBM Watson, to convert it into an openEHR or HL7 standard.	[113]
Knowledge Graphs	Knowledge graphs are used for the interoperability of biomedical data.	[37]

The methods used for interoperability include NLP, data mapping and transformation, data quality assessment, predictive analytics, and anomaly detection. They are used to promote one or more of these: Standardization of data, using application programming interface (API), using middleware and frameworks such as the Da Vinci project, and health information exchanges (HIEs). While effective, NLP techniques are very resource intensive. Data mapping and transformations can be very narrow in application. Data quality assessments can be used to compare inconsistencies but require constant updates and maintenance. Predictive analytics can help improve care coordination and resource allocation, but this is also effective in a narrow range of situations. Anomaly detection can identify unusual or unexpected patterns in healthcare data, potentially flagging issues that may need to be addressed, but can suffer from alert fatigue if the sensitivity is too high. However, it requires certain contextual information to be more effective.

3. Data Fusion: A physiological event can be observed with the help of various sensors, each sensing a unique aspect of the physiological event. The system has to fuse or combine information from different sensing elements for a holistic understanding of the event. This is done at multiple levels. In industry 4.0, healthcare systems, these sensing elements are spread across time and space (wearable sensors, ambulances, and hospitals). Fusing information from multiple sensors provides a more holistic picture of healthcare, including detection, phenotyping, disease progression, and other related data-powered solutions. Ref. [114] exhibits a combination of different layers of data fusion in connected healthcare, from individual sensors to detect medical events, to a network of connected devices, and finally, fusing information amongst various institutions. Ref. [115] displays a sensor fusion model between communication systems. Ref. [116] defines different levels of data fusion. These include signal level fusion, feature level fusion, and decision level fusion. Kalman Filtering is a popular statistics method for signal level fusion and is widely used in biomedical sensor networks. Weighted averages are also widely used to penalize sensors with more

noise in a sensor network [117–119]. Particle filtering, amongst various other variants, is also used extensively for signal level fusion in sensor networks in healthcare [120]. Ref. [121] uses temporal evidence theory for signal level fusion for activity recognition. Feature level fusion means each sensing element's features are calculated and fused. Ref. [122] calculates a linear combination of features to obtain a new feature. Ref. [123] is a weakly supervised program for feature-level fusion. Decision level fusion is a way to fuse decisions based on different information streams. There exist many such systems in the context of healthcare [124,125]. The critical issue in all these is developing a plastic nature of fusion techniques. A plastic fusion technique would be flexible to change with the emerging problem because different features or data may have other significance for each model.

There are several key limitations to data harmonization standards for electronic medical records for example:

- Complexity—Data harmonization standards can be complex and may require significant resources to implement and maintain.
- Limited adoption—Not all electronic medical record systems may adopt the same data harmonization standards, which can limit the ability to exchange data between systems.
- Changing standards—Data standards can change over time, which can make it difficult to maintain compatibility with other systems.
- Privacy and security concerns—The exchange of patient data between systems can raise concerns about privacy and security. Careful measures must be taken to ensure that patient data are protected when they are shared between systems.
- Cost—Implementing and maintaining data harmonization standards can be expensive, particularly for smaller healthcare organizations.
- Intended use—some coding is designed for a different reason than it is used for, e.g., reimbursement versus treatment.

3.5. Decision Systems

The nature of decision systems is specific to the problem they deal with. One axis along which the decision systems can be classified is the medical problems they solve, which include data quality, phenotyping, medication adherence, graph representation of data, detection of disease, and mortality prediction. One more axis along which the decision systems can be classified is the nature of algorithms, natural language processing, time series analysis, and graph neural networks.

1. Data Quality: The quality of the data acquired in healthcare is essential for the credibility of the predicted outcomes. Data quality issues are hard to identify in data with varying structures, shapes, dimensions, and sources. The dimensions of data quality, as elaborated by [126], are completeness (whether the relevant information is present), correctness (are the data correct), concordance (are they relatable to other data sources), plausibility (is any element in the EHRs making sense in the presence of other evidence), and currency (meaning how old are the data). These solutions will help to identify data quality issues, log them, encode them in metadata for datasets, help develop exclusion criteria of data based on its quality, and record the number of such problems. Ref. [127] is one such work that creates a framework to carry out all the tasks and uses probabilistic models to detect temporal stability and plausibility in biomedical data. It employs probabilistic change detection using Jensen–Shannon distance principles of statistical control of posterior beta distribution. Ref. [128] uses probability distribution distance to the same end. Ref. [129] is a measure of completeness by flagging incomplete data sources using the Delphi method. It also measures the same DQ dimension using patterns in the number of patients and compares them. Ref. [130] considers the data quality of radio frequency identification (RFID) in nine phases within healthcare systems.

2. Phenotypes: Phenotypes are the combination of an individual's observable disease traits. The data from the electronic health record are a set of data points related to interventions and the change in the states measured in lab tests. The data help align heterogeneous disease progression into temporal phenotypes. This allows data science techniques to find the relation between disease, symptoms, and interventions. These are also linked to mortality prediction, disease progression, and observation of medically complex phenotypes. Most temporal phenotype identification methods deploy clustering techniques. Phenotypes are also used to identify rare diseases [131–133]. These methods are rule-based [133] and graph-theory-based [134].

One of the critical challenges in AI-based phenotype is the representation of data. The data are being presented to domain experts, but developing a metric that identifies the visual tools' efficacy to represent the temporal phenotypes is worthwhile. For example, in encoding information in the edges and nodes of a graph, silhouette diagrams [135] are very different in richness and application compared to graphs. Graph theory is widely used in these systems as it is very suitable for the relational nature of phenotypes. One key concept is called category theory, which is a directed graph. The data used in temporal phenotypes are time-based (hence, directed), comprising nodes and morphisms. It is different from other graph representations as the morphisms encode the information of the mappings [136]. Very little focused work in this domain comes from EMRs.

3. Deidentification: De-identification of electronic medical records in an automatic manner is an active area of research where blockchain has recently been widely used [137,138]. Ref. [139] compares deep learning, rule-based systems, and shallow learning for de-identifying EMRs and argues that stacked learning is the most efficient ensemble technique. Ref. [140] deploys self-attention networks and stacked recurrent neural networks to de-identify the medical records. The main de-identification methods are neural networks, blockchain technology, and rule-based systems [140]. Some Internet of Medical Things (IoMT) schemes uses IoT protocols to preserve privacy while ensuring that critical information is relayed to the relevant stakeholder [141].

Challenges in this field remain the interplay of structured, unstructured, and semi-structured data. These data come from various sources and categories and, in the case of categorical features with other features, must be collated before solutions can be designed.

4. Adherence: Adherence to suggested and prescribed medical regimens is a crucial component of healthcare. Healthcare is an integrated process; hence, adherence is monitored by different sensing and AI techniques to ensure the efficacy of the interventions. The following Table 6 represents the various AI methods used to this end.

Table 6. Recent Works: AI in Adherence.

Name	Summary	Application	References
Conversational Robot	Chatbot used for drug adherence	Drug Adherence	[142,143]
Ethics	Deliberates over the ethical questions arising from the usage of AI in Norm Adherence	Ethics	[144]
Lifestyle Modification	It uses a web app to help monitor adherence, lifestyle modifications, for Example, in the case of cancer.	Drug Adherence	[145]
Medication Adherence	It uses machine learning to perform binary classification of the medication adherence for Parkinson's disease patients.	Remote Monitoring	[146]
Excercise Adherence	Uses machine learning models to estimate likelihood to adhere to a physical exercise regimen using accelerators and other data sources.	Predictive healthcare	[147]

Table 6. *Cont.*

Name	Summary	Application	References
Medication Adherence	Uses machine learning models to identify the likelihood of non-adherence to medication from electronic health records	Predictive healthcare	[148]
Medication Adherence	Uses data from wearable sensors to measure drug adherence for a specific cause.	Remote Monitoring	[149]
Medication Adherence	Uses cloud-based applications for medication adherence in home hospitalizations	Remote Monitoring	[150]

The key challenge in this domain is access to relevant data as the disease progresses. Here, the importance of different features coming from the same sensors and additional sensors can change as the condition changes its phase.

5. Diagnosis and mortality prediction: Disease prediction can help speed up the process of health care and increase the prediction accuracy, leading to the correct treatment being administered earlier. In the case of critical systems, the idea of mortality prediction and their interplay with demographic information and phenotype can help save lives. It can also help in understanding the progression of the disease and can direct healthcare resources in the right direction. Ref. [151] contains a process for disease prediction using electronic health records. It uses convolutional neural networks (CNN) to this end. Ref. [152] uses hybrid machine learning techniques to predict cardiovascular diseases. It uses a combination of random forest and linear classification models. Ref. [153] develops a naive Bayes analytic model for disease prediction using electronic health records.

Machine learning has been used to predict mortality for some time [154]. It has significant implications for different phenotypes [155,156]. These algorithms are used widely in brain injuries [157,158].

The critical challenges in disease and mortality prediction are the development of explainable machine learning models. As these models make predictions, they need to be explainable and validated as accurate for each patient prediction. Another crucial issue in this domain is the development of ethical frameworks to enable them to be deployed in the real world. Moral dilemmas such as those explained in [159] for self-driving cars should be identified.

3.6. Explainability

The literature uses many exciting techniques in time-series, EHR, and graph-based data. These techniques include feature significance and their interplay-based methods. Deep learning important features, or DeepLIFT, is widely used to this end [160,161] as it combines the importance of a feature as it passes through the layers of the neural network. local interpretable model agnostic explanation, or LIME, introduced in [162], is also widely used in electronic health records [163,164].

Attention mechanisms find the relevant neurons or the dataset components that are the most pertinent information needed for classification. DeepSOFA, DeepHINT, and Grad-CAM are such systems [165–167].

Least absolute shrinkage and selection operator (LASSO) is an explainability technique that uses dimensionality reduction techniques to explain the outcomes of a neural network. They are also used to describe healthcare outcomes [168].

Some explainability techniques draw the rules from the networks, and these systems are also applied in healthcare [169,170].

Deep Taylor decomposition is one such explainability technique used in these systems [171]. Shapley values are also used in such scenarios. The key challenges in developing these systems for graph neural networks are primarily encountered when this method is used for phenotyping.

There is currently a lack of standardized evaluation methods for interpretability techniques, making it difficult to compare and contrast the effectiveness of different approaches. Clinical relevance at present is limited to the identification of which traditional inputs are significant.

3.7. Levels of Automation

Levels of automation for the said topic are discussed as follows:

1. Human Only: Here, there is no AI involved, for example, the calculation of muscle atrophy using electromyogram (EMG) signals [172]. This process, however, involves the signal processing techniques for the representation of data.
2. Shadow Mode: In shadow mode, the data generated by the interaction of the medical practitioner and other sources are logged, and the data are labeled using the judgment of a qualified physician. These data are used to train a machine learning or an optimization algorithm. One such system developed by the ICL team is a reinforcement learning framework optimizing interventions retrospectively that allows a regulatory compliant pathway to clinical testing. This technique is used for sepsis treatment in the ICU [54].
3. AI Assistant: This level of decision making assistance provides the physician with suggestions. Some systems use these to detect cancers; for example, one such system uses biomedical images and structured data to detect hepatocellular carcinoma in the AI assistant model [173].
4. Partial Solutions: Based on the data, the AI comes up with a diagnosis independently, but needs a physician's input.
5. Full Automation: All the tasks in healthcare are provided by AI alone.

4. Conclusions

This review presents a paradigm of the application of AI in times-series and graph-based healthcare data that is driven by translational medicine. It looks at the complete pipeline, starting from data collection, harmonization, and quality dimensions. The decision systems are deliberated over, including various kinds of phenotyping, mortality detection, and other methods. We looked at the components related to the data, classifying them into multiple axes. Recent advances and state of the art technology in the various lexemes of the paradigms found were also reviewed.

Data can be classified along multiple axes, including structure, source, and dimension. Most healthcare data are unstructured, which has been used in conjunction with structured data to predict healthcare outcomes. Data preprocessing techniques can help combine different types of data, denoising and harmonizing to increase the reusability.

Another issue that needs to be tackled in data collection and preprocessing is interoperability of various devices and sensors, and this review has elaborated on different interoperability methods.

There are issues where collected data are fed to different decision systems. This part of the pipeline was discussed by this review, especially where the graph-based solutions, such as temporal phenotyping used to help identify risks for various morbidities and help cluster disease presentation into various groups, are concerned. The most recent works and reviewed literature focus more on applying these solutions in the real world. This application becomes easier when the AI process making predictions can be explained, hence, different explainability and interpretability techniques are compared here while highlighting the lack of standard metrics of evaluation for such methods. The validation of accuracy for each individual patient is an open area of research.

Based on the advances mentioned in this review, any future review may include the identification of ethical dilemmas in healthcare interventions and personalized healthcare: continuous healthcare monitoring and better intervention methods. Clinical use of ambulatory data continues to be a challenge for traditional medical practice. The debate between generalizable AI models with the required precision to achieve individual specific health

outcomes will likely continue. This work can influence the current healthcare system in a positive manner; however, a way of combining these issues can first develop individual specific models, then explain them using explainability techniques, and then cluster them for general exploratory studies.

There are many challenges associated with healthcare data collection for the so-called disease X [174,175]. The more evolved diseases can be stopped from progressing in real-time using multi-omics-profiling and outlier detection [176]. Another challenge in dealing with data derived from time-based sensor data is the integration of advancements in real-time systems. To this end, translational medicine is already defining some solutions. Another major challenge is generating data for groups for which these data are unavailable using generative AI.

Author Contributions: Conceptualization, H.A., I.K.N., B.K.R., C.C. and D.W.; methodology, I.K.N., B.K.R. and S.M.; software, H.A., I.K.N., B.K.R. and S.M.; validation, I.K.N., B.K.R. and S.M.; formal analysis, H.A., I.K.N., B.K.R. and S.M.; investigation, H.A.; resources, I.K.N. and D.W.; data curation, H.A.; writing—original draft preparation, H.A.; writing—review and editing, H.A., I.K.N., B.K.R., C.C., S.M. and D.W.; visualization, H.A.; supervision, I.K.N., B.K.R., S.M. and D.W.; project administration, D.W.; funding acquisition, I.K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the New Zealand College of Chiropractic student scholarship, funding number NZCC 20126384.

Institutional Review Board Statement: Ethical approval was not required for publication review.

Informed Consent Statement: Informed consent was not required for publication review.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hartl, D.; de Luca, V.; Kostikova, A.; Laramie, J.; Kennedy, S.; Ferrero, E.; Siegel, R.; Fink, M.; Ahmed, S.; Millholland, J.; et al. Translational precision medicine: An industry perspective. *J. Transl. Med.* **2021**, *19*, 245. [CrossRef] [PubMed]
2. Jordan, L. The problem with Big Data in Translational Medicine. A review of where we've been and the possibilities ahead. *Appl. Transl. Genom.* **2015**, *6*, 3–6. [CrossRef] [PubMed]
3. Ewusie, J.; Soobiah, C.; Blondal, E.; Beyene, J.; Thabane, L.; Hamid, J.S. Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review. *J. Multidiscip. Health* **2020**, *13*, 411–423. [CrossRef]
4. Ahmad, M.A.; Eckert, C.; Teredesai, A. Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 29 August–1 September 2018.
5. Baum, E.B. On the capabilities of multilayer perceptrons. *J. Complex.* **1988**, *4*, 193–215. [CrossRef]
6. Paganelli, A.I.; Mondéjar, A.G.; da Silva, A.C.; Silva-Calpa, G.; Teixeira, M.F.; Carvalho, F.; Raposo, A.; Endler, M. Real-time data analysis in health monitoring systems: A comprehensive systematic literature review. *J. Biomed. Inform.* **2022**, *127*, 104009. [CrossRef] [PubMed]
7. Chen, J.X. The Evolution of Computing: AlphaGo. *Comput. Sci. Eng.* **2016**, *18*, 4–7. [CrossRef]
8. Singh, S.P.; Wang, L.; Gupta, S.; Goli, H.; Padmanabhan, P.; Gulyás, B. 3D Deep Learning on Medical Images: A Review. *Sensors* **2020**, *20*, 5097. [CrossRef]
9. Taghanaki, S.A.; Abhishek, K.; Cohen, J.P.; Cohen-Adad, J.; Hamarneh, G. Deep semantic segmentation of natural and medical images: A review. *Artif. Intell. Rev.* **2020**, *54*, 137–178. [CrossRef]
10. Kumar, M.; Mishra, S.K. A comprehensive review on nature inspired neural network based adaptive filter for eliminating noise in medical images. *Curr. Med. Imaging* **2020**, *16*, 278. [CrossRef] [PubMed]
11. Pavlič, J.; Tomažič, T.; Kožuh, I. The impact of emerging technology influences product placement effectiveness: A scoping study from interactive marketing perspective. *J. Res. Interact. Mark.* **2021**, *16*, 551–568. [CrossRef]
12. Davy, C.; Bleasel, J.; Liu, H.; Tchan, M.; Ponniah, S.; Brown, A. Effectiveness of chronic care models: Opportunities for improving healthcare practice and health outcomes: A systematic review. *BMC Health Serv. Res.* **2015**, *15*, 1–11. [CrossRef] [PubMed]
13. Chen, I.Y.; Joshi, S.; Ghassemi, M.; Ranganath, R. Probabilistic Machine Learning for Healthcare. *Annu. Rev. Biomed. Data Sci.* **2021**, *4*, 393–415. [CrossRef]

14. Wang, F.; Cui, P.; Pei, J.; Song, Y.; Zang, C. Recent Advances on Graph Analytics and Its Applications in Healthcare. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 6–10 July 2020.
15. Gu, D.; Li, T.; Wang, X.; Yang, X.; Yu, Z. Visualizing the intellectual structure and evolution of electronic health and telemedicine research. *Int. J. Med. Inform.* **2019**, *130*, 103947. [CrossRef] [PubMed]
16. Tran, B.X.; Nghiem, S.; Sahin, O.; Vu, T.M.; Ha, G.H.; Vu, G.T.; Pham, H.Q.; Do, H.T.; Latkin, C.; Tam, W.; et al. Modeling Research Topics for Artificial Intelligence Applications in Medicine: Latent Dirichlet Allocation Application Study. *J. Med. Internet Res.* **2019**, *21*, e15511. [CrossRef]
17. Adnan, K.; Akbar, R.; Khor, S.W.; Ali, A.B.A. Role and Challenges of Unstructured Big Data in Healthcare. In *Data Management, Analytics and Innovation*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 301–323.
18. Krishna, S.; Boren, S.; Balas, E.A. Healthcare via Cell Phones: A Systematic Review. *Telemed. e-Health* **2009**, *15*, 231–240. [CrossRef]
19. Bellamy, D.; Celi, L.; Beam, A. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv* **2020**, arXiv:2010.01149.
20. McCarthy, J.J. Optimal paradigms. In *Paradigms in Phonological Theory*; Linguistics Department Faculty Publication: Amherst, MA, USA, 2005; p. 55.
21. Adithan, C. Principles of translational science in medicine: From bench to bedside. *Indian J. Med. Res.* **2017**, *145*, 408–409. [CrossRef]
22. Liu, B.; Li, J.; Chen, C.; Tan, W.; Chen, Q.; Zhou, M. Efficient Motif Discovery for Large-Scale Time Series in Healthcare. *IEEE Trans. Ind. Informatics* **2015**, *11*, 583–590. [CrossRef]
23. Balasubramanian, A.; Wang, J.; Prabhakaran, B. Discovering multidimensional motifs in physiological signals for personalized healthcare. *IEEE J. Sel. Top. Signal Process.* **2016**, *10*, 832–841. [CrossRef]
24. Pereira, J.; Silveira, M. Learning representations from healthcare time series data for unsupervised anomaly detection. In Proceedings of the 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), Kyoto, Japan, 27 February–02 March 2019.
25. Kaushik, S.; Choudhury, A.; Sheron, P.K.; Dasgupta, N.; Natarajan, S.; Pickett, L.A.; Dutt, V. AI in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures. *Front. Big Data* **2020**, *3*, 4. [CrossRef] [PubMed]
26. Maweu, B.M.; Shamsuddin, R.; Dakshit, S.; Prabhakaran, B. Generating Healthcare Time Series Data for Improving Diagnostic Accuracy of Deep Neural Networks. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 2508715. [CrossRef]
27. Harutyunyan, H.; Khachatrian, H.; Kale, D.C.; Ver Steeg, G.; Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **2019**, *6*, 96. [CrossRef] [PubMed]
28. Lipton, Z.C.; Kale, D.; Wetzel, R. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In Proceedings of the 1st Machine Learning for Healthcare Conference, Los Angeles, CA, USA, 9–20 August 2016.
29. Whitney, V.K.M. Relational data management implementation techniques. In Proceedings of the ACM SIGFIDET (now SIGMOD) Workshop on Data description, Access and Control, Ann Arbor, MI, USA, 1–5 May 1974.
30. Zhang, Y.; Sheng, M.; Zhou, R.; Wang, Y.; Han, G.; Zhang, H.; Xing, C.; Dong, J. Hkgb: An inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. *Inf. Process. Manag.* **2020**, *57*, 102324. [CrossRef]
31. Choi, E.; Bahadori, M.T.; Song, L.; Stewart, W.F.; Sun, J. GRAM: Graph-based attention model for healthcare representation learning. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, Halifax, NS, Canada, 13–17 August 2017.
32. Liu, C.; Wang, F.; Hu, J.; Xiong, H. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, NSW, Australia, 10–13 August 2015.
33. Sood, S.K.; Mahajan, I. A Fog-Based Healthcare Framework for Chikungunya. *IEEE Internet Things J.* **2017**, *5*, 794–801. [CrossRef]
34. Wu, T.; Wang, Y.; Wang, Y.; Zhao, E.; Yuan, Y. Leveraging graph-based hierarchical medical entity embedding for healthcare applications. *Sci. Rep.* **2021**, *11*, 5858. [CrossRef] [PubMed]
35. Winter, A.; Brigl, B.; Wendt, T. Modeling hospital information systems (part 1): The revised three-layer graph-based meta model 3LGM2. *Methods Inf. Med.* **2003**, *42*, 544–551. [PubMed]
36. Sharma, N.; Bhatt, R. Privacy Preservation in WSN for Healthcare Application. *Procedia Comput. Sci.* **2018**, *132*, 1243–1252. [CrossRef]
37. Malik, K.M.; Krishnamurthy, M.; Alobaidi, M.; Hussain, M.; Alam, F.; Malik, G. Automated domain-specific healthcare knowledge graph curation framework: Subarachnoid hemorrhage as phenotype. *Expert Syst. Appl.* **2019**, *145*, 113120. [CrossRef]
38. Kalamaras, I.; Glykos, K.; Megalooikonomou, V.; Votis, K.; Tzovaras, D. Graph-based visualization of sensitive medical data. *Multimed. Tools Appl.* **2021**, *81*, 209–236. [CrossRef]
39. Zhang, Q.; Zhou, J.; Zhang, B. Graph Based Multichannel Feature Fusion for Wrist Pulse Diagnosis. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 3732–3743. [CrossRef] [PubMed]
40. Finlayson, S.G.; LePendu, P.; Shah, N.H. Building the graph of medicine from millions of clinical narratives. *Sci. Data* **2014**, *1*, 140032. [CrossRef]

41. Yue, X.; Wang, Z.; Huang, J.; Parthasarathy, S.; Moosavinasab, S.; Huang, Y.; Lin, S.M.; Zhang, W.; Zhang, P.; Sun, H. Graph embedding on biomedical networks: Methods, applications and evaluations. *Bioinformatics* **2019**, *36*, 1241–1251. [CrossRef] [PubMed]
42. Zhang, W.; Chen, Y.; Li, D.; Yue, X. Manifold regularized matrix factorization for drug-drug interaction prediction. *J. Biomed. Inform.* **2018**, *88*, 90–97. [CrossRef]
43. Kulmanov, M.; Khan, M.A.; Hoehndorf, R. DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **2017**, *34*, 660–668. [CrossRef]
44. Chen, L.; Li, X.; Sheng, Q.Z.; Peng, W.-C.; Bennett, J.; Hu, H.-Y.; Huang, N. Mining Health Examination Records—A Graph-Based Approach. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 2423–2437. [CrossRef]
45. Kaur, K.; Rani, R. Managing Data in Healthcare Information Systems: Many Models, One Solution. *Computer* **2015**, *48*, 52–59. [CrossRef]
46. Thews, O.; Rohrbach, C.; Sergl, M.; Pommerening, K.; Müller, R. A Graph-Grammar Approach to Represent Causal, Temporal and Other Contexts in an Oncological Patient Record. *Methods Inf. Med.* **1996**, *35*, 127–141. [CrossRef]
47. Liu, Y.; Song, Z.; Xu, X.; Rafique, W.; Zhang, X.; Shen, J.; Khosravi, M.R.; Qi, L. Bidirectional GRU networks-based next POI category prediction for healthcare. *Int. J. Intell. Syst.* **2021**, *37*, 4020–4040. [CrossRef]
48. Rodeheaver, N.; Kim, H.; Herbert, R.; Seo, H.; Yeo, W.H. Breathable, Wireless, Thin-Film Wearable Biopatch Using Noise-Reduction Mechanisms. *ACS Appl. Electron. Mater.* **2022**, *4*, 503–512. [CrossRef]
49. Yang, J.; Soltan, A.A.S.; Clifton, D.A. Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. *NPJ Digit. Med.* **2022**, *5*, 69. [CrossRef] [PubMed]
50. Palanisamy, V.; Thirunavukarasu, R. Implications of big data analytics in developing healthcare frameworks—A review. *J. King Saud Univ. Comput. Inf. Sci.* **2019**, *31*, 415–425. [CrossRef]
51. Moore, W.; Frye, S. Review of HIPAA, part 1: History, protected health information, and privacy and security rules. *J. Nucl. Med. Technol.* **2019**, *47*, 269–272. [CrossRef]
52. Zhang, Z.; Yan, C.; Lasko, T.A.; Sun, J.; Malin, B. SynTEG: A framework for temporal structured electronic health data simulation. *J. Am. Med. Inform. Assoc.* **2020**, *28*, 596–604. [CrossRef]
53. Abedi, M.; Hempel, L.; Sadeghi, S.; Kirsten, T. GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Appl. Sci.* **2022**, *12*, 7075. [CrossRef]
54. Li, L.; Albert-Smet, I.; Faisal, A. Optimizing medical treatment for sepsis in intensive care: From reinforcement learning to pre-trial evaluation. *arXiv* **2020**, arXiv:2003.06474.
55. Klompas, M.; Kulldorff, M.; Vilk, Y.; Bialek, S.R.; Harpaz, R. Herpes Zoster and Postherpetic Neuralgia Surveillance Using Structured Electronic Data. *Mayo Clin. Proc.* **2011**, *86*, 1146–1153. [CrossRef] [PubMed]
56. Aminifar, A.; Lamo, Y.; Pun, K.I.; Rabbi, F. *A Practical Methodology for Anonymization of Structured Health Data*; Linköping University Electronic Press: Linköping, Sweden, 2019.
57. Kanwal, T.; Anjum, A.; Khan, A. Privacy preservation in e-health cloud: Taxonomy, privacy requirements, feasibility analysis, and opportunities. *Clust. Comput.* **2020**, *24*, 293–317. [CrossRef]
58. Xu, R.; Baracaldo, N.; Zhou, Y.; Anwar, A.; Ludwig, H. Hybridalpha: An efficient approach for privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, London, UK, 15 November 2019.
59. Wibawa, F.; Catak, F.O.; Kuzlu, M.; Sarp, S.; Cali, U. Homomorphic Encryption and Federated Learning based Privacy-Preserving CNN Training: COVID-19 Detection Use-Case. In Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference, Barcelona, Spain, 15–16 June 2022.
60. Wright, A.; McEvoy, D.S.; Aaron, S.; McCoy, A.; Amato, M.G.; Kim, H.; Ai, A.; Cimino, J.J.; Desai, B.R.; El-Kareh, R.; et al. Structured override reasons for drug-drug interaction alerts in electronic health records. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 934–942. [CrossRef] [PubMed]
61. Vest, J.R.; Grannis, S.J.; Haut, D.P.; Halverson, P.K.; Menachemi, N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int. J. Med. Inform.* **2017**, *107*, 101–106. [CrossRef] [PubMed]
62. Wang, W.; Kiik, M.; Peek, N.; Curcin, V.; Marshall, I.J.; Rudd, A.G.; Wang, Y.; Douiri, A.; Wolfe, C.D.; Bray, B. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS ONE* **2020**, *15*, e0234722.
63. Richter, A.N.; Khoshgoftaar, T.M. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artif. Intell. Med.* **2018**, *90*, 1–14. [CrossRef] [PubMed]
64. Therneau, T.M.; Grambsch, P.M. *Modeling Survival Data: Extending the Cox Model*; Springer: New York, NY, USA, 2000; pp. 39–77.
65. Gopinath, D.; Agrawal, M.; Murray, L.; Horng, S.; Karger, D.; Sontag, D. Fast, Structured Clinical Documentation via Contextual Autocomplete. In Proceedings of the 5th Machine Learning for Healthcare Conference, Virtual, 7–8 August 2020.
66. Shao, Y.; Zeng, Q.T.; Chen, K.K.; Shutes-David, A.; Thielke, S.M.; Tsuang, D.W. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 128. [CrossRef] [PubMed]
67. Banda, J.M.; Seneviratne, M.; Hernandez-Boussard, T.; Shah, N.H. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu. Rev. Biomed. Data Sci.* **2018**, *1*, 53–68. [CrossRef] [PubMed]
68. Sung, S.-F.; Lin, C.-Y.; Hu, Y.-H. EMR-Based Phenotyping of Ischemic Stroke Using Supervised Machine Learning and Text Mining Techniques. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2922–2931. [CrossRef] [PubMed]

69. Boxwala, A.; Kim, J.; Grillo, J.M.; Ohno-Machado, L. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 498–505. [CrossRef] [PubMed]
70. Lantz, E. Machine Learning for Risk Prediction and Privacy in Electronic Health Records. Ph.D. Thesis, The University of Wisconsin-Madison, Madison, WI, USA, 2016.
71. Kim, S.; Lee, H.; Chung, Y.D. Privacy-preserving data cube for electronic medical records: An experimental evaluation. *Int. J. Med. Inform.* **2017**, *97*, 33–42. [CrossRef] [PubMed]
72. Marble, H.D.; Huang, R.; Dudgeon, S.N.; Lowe, A.; Herrmann, M.D.; Blakely, S.; Leavitt, M.O.; Isaacs, M.; Hanna, M.G.; Sharma, A.; et al. A Regulatory Science Initiative to Harmonize and Standardize Digital Pathology and Machine Learning Processes to Speed up Clinical Innovation to Patients. *J. Pathol. Inform.* **2020**, *11*, 22. [CrossRef] [PubMed]
73. Guan, J.; Li, R.; Yu, S.; Zhang, X. A method for generating synthetic electronic medical record text. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *23*, 173–182. [CrossRef]
74. Chin-Cheong, K.; Sutter, T.M.; Vogt, J.E. Generation of heterogeneous synthetic electronic health records using GANs. In Proceedings of the Workshop on Machine Learning For Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
75. Walonoski, J.A.; Kramer, M.; Nichols, J.; Quina, A.; Moesel, C.; Hall, D.; Duffett, C.; Dube, K.; Gallagher, T.; McLachlan, S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inform. Assoc.* **2017**, *25*, 230–238. [CrossRef]
76. Chen, I.Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; Ghassemi, M. Ethical Machine Learning in Healthcare. *Annu. Rev. Biomed. Data Sci.* **2020**, *4*, 123–144. [CrossRef] [PubMed]
77. Aggarwal, A.; Garhwal, S.; Kumar, A. HEDEA: A Python Tool for Extracting and Analysing Semi-structured Information from Medical Records. *Health Inform. Res.* **2018**, *24*, 148–153. [CrossRef] [PubMed]
78. Makarova, E.; Lagerev, D. Methodology for Preprocessing Semi-Structured Data for Making Managerial Decisions in the Healthcare. In Proceedings of the InCEUR Workshop Proceedings of the 30th International Conference on Computer Graphics and Machine Vision, Saint Petersburg, Russia, 22–25 September 2020.
79. Hong, N.; Wen, A.; Stone, D.J.; Tsuji, S.; Kingsbury, P.R.; Rasmussen, L.V.; Pacheco, J.A.; Adekkanattu, P.; Wang, F.; Luo, Y.; et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J. Biomed. Inform.* **2019**, *99*, 103310. [CrossRef]
80. Batra, S.; Sachdeva, S. Organizing standardized electronic healthcare records data for mining. *Health Policy Technol.* **2016**, *5*, 226–242. [CrossRef]
81. Yuan, J.; Holtz, C.; Smith, T.H.; Luo, J. Autism spectrum disorder detection from semi-structured and unstructured medical data. *EURASIP J. Bioinform. Syst. Biol.* **2016**, *2017*, 3. [CrossRef]
82. Miled, Z.B.; Haas, K.; Black, C.M.; Khandker, R.K.; Chandrasekaran, V.; Lipton, R.; Boustani, M.A. Predicting dementia with routine care EMR data. *Artif. Intell. Med.* **2020**, *102*, 101771. [CrossRef]
83. Geraci, J.; Wilansky, P.; de Luca, V.; Roy, A.; Kennedy, J.L.; Strauss, J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evid.-Based Ment. Health* **2017**, *20*, 83–87. [CrossRef] [PubMed]
84. Goh, K.H.; Wang, L.; Yeow, A.Y.K.; Poh, H.; Li, K.; Yeow, J.J.L.; Tan, G.Y.H. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat. Commun.* **2021**, *12*, 1–10. [CrossRef] [PubMed]
85. Zuo, Z.; Li, J.; Xu, H.; Al Moubayed, N. Curvature-based feature selection with application in classifying electronic health records. *Technol. Forecast. Soc. Chang.* **2021**, *173*, 12112. [CrossRef]
86. Xu, Z.; Chou, J.; Zhang, X.S.; Luo, Y.; Isakova, T.; Adekkanattu, P.; Ancker, J.S.; Jiang, G.; Kiefer, R.C.; Pacheco, J.A.; et al. Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks. *J. Biomed. Inform.* **2020**, *102*, 103361. [CrossRef] [PubMed]
87. Zhao, M.; Havrilla, J.; Peng, J.; Drye, M.; Fecher, M.; Guthrie, W.; Tunc, B.; Schultz, R.; Wang, K.; Zhou, Y. Development of a phenotype ontology for autism spectrum disorder by natural language processing on electronic health records. *J. Neurodev. Disord.* **2022**, *14*, 32. [CrossRef] [PubMed]
88. Song, B.; Feng, Y.; Li, X.; Sun, Z.; Yang, Y. Un-apriori: A novel association rule mining algorithm for unstructured EMRs. In Proceedings of the IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), Dalian, China, 12–15 October 2017.
89. Kim, J.-C.; Chung, K. Associative Feature Information Extraction Using Text Mining from Health Big Data. *Wirel. Pers. Commun.* **2018**, *105*, 691–707. [CrossRef]
90. Boustani, M.; Perkins, A.J.; Khandker, R.K.; Duong, S.; Dexter, P.R.; Lipton, R.; Black, C.M.; Chandrasekaran, V.; Solid, C.A.; Monahan, P. Passive digital signature for early identification of Alzheimer's disease and related dementia. *J. Am. Geriatr. Soc.* **2020**, *68*, 511–518. [CrossRef]
91. Chung, K.; Yoo, H.; Choe, D.-E. Ambient context-based modeling for health risk assessment using deep neural network. *J. Ambient. Intell. Humaniz. Comput.* **2018**, *11*, 1387–1395. [CrossRef]
92. Ling, A.Y.; Kurian, A.W.; Caswell-Jin, J.; Sledge, G.W.; Shah, N.H.; Tamang, S.R. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. *JAMIA Open* **2019**, *2*, 528–553. [CrossRef] [PubMed]

93. Wallace, B.C.; Kuiper, J.; Sharma, A.; Zhu, M.B.; Marshall, I.J. Extracting PICO Sentences from Clinical Trial Reports using Supervised Distant Supervision. *J. Mach. Learn. Res.* **2016**, *17*, 4572–4596.
94. Lin, F.P.-Y.; Pokorny, A.; Teng, C.; Epstein, R.J. TEPAPA: A novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. *Sci. Rep.* **2017**, *7*, 6918. [CrossRef] [PubMed]
95. Bjarnadottir, R.I.; Lucero, R.J. What Can We Learn about Fall Risk Factors from EHR Nursing Notes? A Text Mining Study. *eGEMs* **2018**, *6*, 21. [CrossRef]
96. Wang, L.; Lou, Z.; Jiang, K.; Shen, G. Bio-Multifunctional Smart Wearable Sensors for Medical Devices. *Adv. Intell. Syst.* **2019**, *1*, 1900040. [CrossRef]
97. Sempionatto, J.R.; Montiel, V.R.-V.; Vargas, E.; Teymourian, H.; Wang, J. Wearable and Mobile Sensors for Personalized Nutrition. *ACS Sens.* **2021**, *6*, 1745–1760. [CrossRef]
98. Chai, P.R.; Goodman, G.; Bustamante, M.; Mendez, L.; Mohamed, Y.; Mayer, K.H.; Boyer, E.W.; Rosen, R.K.; O'Cleirigh, C. Design and Delivery of Real-Time Adherence Data to Men Who Have Sex with Men Using Antiretroviral Pre-exposure Prophylaxis via an Ingestible Electronic Sensor. *AIDS Behav.* **2020**, *25*, 1661–1674. [CrossRef] [PubMed]
99. Weitschies, W.; Müller, L.; Grimm, M.; Koziolek, M. Ingestible devices for studying the gastrointestinal physiology and their application in oral biopharmaceutics. *Adv. Drug Deliv. Rev.* **2021**, *176*, 113853. [CrossRef]
100. Li, G.; Lian, W.; Qu, H.; Li, Z.; Zhou, Q.; Tian, J. Improving patient care through the development of a 5G-powered smart hospital. *Nat. Med.* **2021**, *27*, 936–937. [CrossRef]
101. Muhammad, G.; Alshehri, F.; Karray, F.; El Saddik, A.; Alsulaiman, M.; Falk, T.H. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Inf. Fusion* **2021**, *76*, 355–375.
102. Lucas, C.; Wong, P.; Klein, J.; Castro, T.B.R.; Silva, J.; Sundaram, M.; Ellingson, M.K.; Mao, T.; Oh, J.E.; Israelow, B.; et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **2020**, *584*, 463–469. [CrossRef] [PubMed]
103. Batra, G.; Aktaa, S.; Wallentin, L.; Maggioni, A.P.; Wilkinson, C.; Casadei, B.; Gale, C.P. Methodology for the development of international clinical data standards for common cardiovascular conditions: European Unified Registries for Heart Care Evaluation and Randomised Trials (EuroHeart). *Eur. Heart J. Qual. Care Clin. Outcomes* **2021**, *2021*, qcab052. [CrossRef] [PubMed]
104. Baxter, S.L.; Lee, A.Y. Gaps in standards for integrating artificial intelligence technologies into ophthalmic practice. *Curr. Opin. Ophthalmol.* **2021**, *32*, 431–438. [CrossRef]
105. American Diabetes Association. Diabetes technology: Standards of medical care in diabetes—2021. *Diabetes Care* **2021**, *44*, S85–S99. [CrossRef] [PubMed]
106. Laleci, G.B.; Dogac, A. A Semantically Enriched Clinical Guideline Model Enabling Deployment in Heterogeneous Healthcare Environments. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *13*, 263–273. [CrossRef]
107. Cartwright, D.J. ICD-9-CM to ICD-10-CM Codes: What? Why? How? *Adv. Wound Care* **2013**, *2*, 588–592. [CrossRef]
108. Dotson, P. *CPT® Codes: What Are They, Why Are They Necessary, and How Are They Developed?* Mary Ann Liebert: New Rochelle, NY, USA, 2013.
109. Gordon, W.J.; Catalini, C. Blockchain Technology for Healthcare: Facilitating the Transition to Patient-Driven Interoperability. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 224–230. [CrossRef] [PubMed]
110. Jabbar, R.; Fetais, N.; Krichen, M.; Barkaoui, K. Blockchain technology for healthcare: Enhancing shared electronic health record interoperability and integrity. In Proceedings of the IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, 2–5 February 2020; pp. 310–317.
111. Pathak, N.; Misra, S.; Mukherjee, A.; Kumar, N. HeDI: Healthcare Device Interoperability for IoT-Based e-Health Platforms. *IEEE Internet Things J.* **2021**, *8*, 16845–16852. [CrossRef]
112. Balakrishna, S.; Thirumaran, M. Semantic Interoperability in IoT and Big Data for Health Care: A Collaborative Approach. In *Handbook of Data Science Approaches for Biomedical Engineering*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 185–220.
113. Joshi, R.; Negi, S.; Sachdeva, S. *Cloud Based Interoperability in Healthcare, in Computational Methods and Data Engineering*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 599–611.
114. Dautov, R.; Distefano, S.; Buyya, R. Hierarchical data fusion for Smart Healthcare. *J. Big Data* **2019**, *6*, 19. [CrossRef]
115. Hall, D.L.; Llinas, J. An introduction to multisensor data fusion. *Proc. IEEE* **1997**, *85*, 6–23. [CrossRef]
116. Lee, H.; Park, K.; Lee, B.; Choi, J.; Elmasri, R. Issues in data fusion for healthcare monitoring. In Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments, Athens, Greece, 16–18 July 2008.
117. Djenouri, D.; Balasingham, I. New QoS and geographical routing in wireless biomedical sensor networks. In Proceedings of the Sixth International Conference on Broadband Communications, Networks, and Systems, Madrid, Spain, 14–16 September 2009.
118. Choi, S.; Han, S.I.; Jung, D.; Hwang, H.J.; Lim, C.; Bae, S.; Park, O.K.; Tschabrunn, C.M.; Lee, M.; Bae, S.Y.; et al. Highly conductive, stretchable and biocompatible Ag–Au core–sheath nanowire composite for wearable and implantable bioelectronics. *Nat. Nanotechnol.* **2018**, *13*, 1048–1056. [CrossRef] [PubMed]
119. Nathan, V.; Jafari, R. Particle Filtering and Sensor Fusion for Robust Heart Rate Monitoring Using Wearable Sensors. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 1834–1846. [CrossRef]
120. Brady, K.; Gwon, Y.; Khorrami, P.; Godoy, E.; Campbell, W.; Dagli, C.; Huang, T.S. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016.

121. McKeever, S.; Ye, J.; Coyle, L.; Bleakley, C.; Dobson, S. Activity recognition using temporal evidence theory. *J. Ambient. Intell. Smart Environ.* **2010**, *2*, 253–269. [CrossRef]
122. Cai, H.; Qu, Z.; Li, Z.; Zhang, Y.; Hu, X.; Hu, B. Feature-level fusion approaches based on multimodal EEG data for depression recognition. *Inf. Fusion* **2020**, *59*, 127–138. [CrossRef]
123. Miao, F.; Liu, Z.-D.; Liu, J.-K.; Wen, B.; He, Q.-Y.; Li, Y. Multi-Sensor Fusion Approach for Cuff-Less Blood Pressure Measurement. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 79–91. [CrossRef] [PubMed]
124. Hossain, M.S.; Muhammad, G. Emotion-Aware Connected Healthcare Big Data Towards 5G. *IEEE Internet Things J.* **2017**, *5*, 2399–2406. [CrossRef]
125. Chen, C.; Jafari, R.; Kehtarnavaz, N. A Real-Time Human Action Recognition System Using Depth and Inertial Sensor Fusion. *IEEE Sens. J.* **2015**, *16*, 773–781. [CrossRef]
126. Weiskopf, N.G.; Weng, C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 144–151. [CrossRef]
127. Fox, F.; Aggarwal, V.R.; Whelton, H.; Johnson, O. A data quality framework for process mining of electronic health record data. In Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, USA, 4–7 June 2018.
128. Sáez, C.; Rodrigues, P.P.; Gama, J.; Robles, M.; Garcia-Gomez, J.M. Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. *Data Min. Knowl. Discov.* **2014**, *29*, 950–975. [CrossRef]
129. Puttkammer, N.; Baseman, J.; Devine, E.; Valles, J.; Hyppolite, N.; Garilus, F.; Honoré, J.; Matheson, A.; Zeliadt, S.; Yuhas, K.; et al. An assessment of data quality in a multi-site electronic medical record system in Haiti. *Int. J. Med. Inform.* **2016**, *86*, 104–116. [CrossRef] [PubMed]
130. Taggart, J.; Liaw, S.-T.; Yu, H. Structured data quality reports to improve EHR data quality. *Int. J. Med. Inform.* **2015**, *84*, 1094–1098. [CrossRef] [PubMed]
131. Li, Q.; Zhao, K.; Bustamante, C.D.; Ma, X.; Wong, W.H. Xrare: A machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Anesthesia Analg.* **2019**, *21*, 2126–2134. [CrossRef]
132. Jia, J.; Wang, R.; An, Z.; Guo, Y.; Ni, X.; Shi, T. RDAD: A Machine Learning System to Support Phenotype-Based Rare Disease Diagnosis. *Front. Genet.* **2018**, *9*, 587. [CrossRef] [PubMed]
133. Morley, K.I.; Wallace, J.; Denaxas, S.C.; Hunter, R.J.; Patel, R.S.; Perel, P.; Shah, A.D.; Timmis, A.D.; Schilling, R.J.; Hemingway, H. Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. *PLoS ONE* **2014**, *9*, e110900. [CrossRef] [PubMed]
134. Ash, J.A.; Rapp, P.R. A quantitative neural network approach to understanding aging phenotypes. *Ageing Res. Rev.* **2014**, *15*, 44–50. [CrossRef]
135. Lee, C.; Rashbass, J.; van der Schaar, M. Outcome-Oriented Deep Temporal Phenotyping of Disease Progression. *IEEE Trans. Biomed. Eng.* **2020**, *68*, 2423–2434. [CrossRef]
136. Tuyéras, R. Category theory for genetics II: Genotype, phenotype and haplotype. *arXiv* **2018**, arXiv:1805.07004.
137. Mayer, A.H.; da Costa, C.; Righi, R. Electronic health records in a blockchain: A systematic review. *Health Inform. J.* **2020**, *26*, 1273–1288. [CrossRef]
138. Shi, S.; He, D.; Li, L.; Kumar, N.; Khan, M.K.; Choo, K.-K.R. Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey. *Comput. Secur.* **2020**, *97*, 101966. [CrossRef]
139. Kim, Y.; Heider, P.; Meystre, S. Ensemble-based Methods to Improve De-identification of Electronic Health Record Narratives. *AMIA Annu. Symp. Proceedings AMIA Symp.* **2018**, *2018*, 663–672.
140. Ahmed, T.; Al Aziz, M.; Mohammed, N. De-identification of electronic health record using neural network. *Sci. Rep.* **2020**, *10*, 18600. [CrossRef]
141. Guan, Z.; Lv, Z.; Du, X.; Wu, L.; Guizani, M. Achieving data utility-privacy tradeoff in Internet of Medical Things: A machine learning approach. *Futur. Gener. Comput. Syst.* **2019**, *98*, 60–68. [CrossRef]
142. Vaidyam, A.N.; Wisniewski, H.; Halamka, J.D.; Kashavan, M.S.; Torous, J.B. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can. J. Psychiatry* **2019**, *64*, 456–464. [CrossRef] [PubMed]
143. Abd-Alrazaq, A.; Safi, Z.; Alajlani, M.; Warren, J.; Househ, M.; Denecke, K. Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. *J. Med. Internet Res.* **2020**, *22*, e18301. [CrossRef] [PubMed]
144. Campbell, J.I.; Eyal, N.; Musiimenta, A.; Haberer, J.E. Ethical Questions in Medical Electronic Adherence Monitoring. *J. Gen. Intern. Med.* **2015**, *31*, 338–342. [CrossRef]
145. Golshahi, J.; Ahmadzadeh, H.; Sadeghi, M.; Mohammadifard, N.; Pourmoghaddas, A. Effect of self-care education on lifestyle modification, medication adherence and blood pressure in hypertensive adults: Randomized controlled clinical trial. *Adv. Biomed. Res.* **2015**, *4*, 204–209.
146. Molugulu, N.; Gubbiyappa, K.S.; Murthy, C.R.V.; Lumae, L.; Mruthyunjaya, A.T. Evaluation of self-reported medication adherence and its associated factors among epilepsy patients in Hospital Kuala Lumpur. *J. Basic Clin. Pharm.* **2016**, *7*, 105–109. [CrossRef] [PubMed]
147. Bavan, L.; Surmacz, K.; Beard, D.; Mellon, S.; Rees, J. Adherence monitoring of rehabilitation exercise with inertial sensors: A clinical validation study. *Gait Posture* **2019**, *70*, 211–217. [CrossRef] [PubMed]

148. Wang, L.; Fan, R.; Zhang, C.; Hong, L.; Zhang, T.; Chen, Y.; Liu, K.; Wang, Z.; Zhong, J. Applying Machine Learning Models to Predict Medication Nonadherence in Crohn's Disease Maintenance Therapy. *Patient Prefer. Adherence* **2020**, *14*, 917–926. [CrossRef] [PubMed]
149. Aldeer, M.; Javanmard, M.; Martin, R.P. A Review of Medication Adherence Monitoring Technologies. *Appl. Syst. Innov.* **2018**, *1*, 14. [CrossRef]
150. Chai, P.R.; Castillo-Mancilla, J.; Buffkin, E.; Darling, C.; Rosen, R.K.; Horvath, K.J.; Boudreaux, E.D.; Robbins, G.K.; Hibberd, P.L.; Boyer, E.W. Utilizing an Ingestible Biosensor to Assess Real-Time Medication Adherence. *J. Med. Toxicol.* **2015**, *11*, 439–444. [CrossRef] [PubMed]
151. Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease Prediction by Machine Learning Over Big Data from Healthcare Communities. *IEEE Access* **2017**, *5*, 8869–8879. [CrossRef]
152. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **2019**, *7*, 81542–81554. [CrossRef]
153. Venkatesh, R.; Balasubramanian, C.; Kaliappan, M. Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique. *J. Med. Syst.* **2019**, *43*, 272. [CrossRef] [PubMed]
154. Cooper, G.F.; Aliferis, C.F.; Ambrosino, R.; Aronis, J.; Buchanan, B.G.; Caruana, R.; Fine, M.J.; Glymour, C.; Gordon, G.; Hanusa, B.H.; et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif. Intell. Med.* **1997**, *9*, 107–138. [CrossRef]
155. Rose, S. Mortality Risk Score Prediction in an Elderly Population Using Machine Learning. *Am. J. Epidemiol.* **2013**, *177*, 443–452. [CrossRef]
156. van Doorn, W.P.; Stassen, P.M.; Borggreve, H.F.; Schalkwijk, M.J.; Stoffers, J.; Bekers, O.; Meex, S.J. A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS ONE* **2021**, *16*, e0245157. [CrossRef] [PubMed]
157. Raj, R.; Luostarinen, T.; Pursiainen, E.; Posti, J.P.; Takala, R.S.K.; Bendel, S.; Konttila, T.; Korja, M. Machine learning-based dynamic mortality prediction after traumatic brain injury. *Sci. Rep.* **2019**, *9*, 17672. [CrossRef]
158. Rau, C.-S.; Kuo, P.-J.; Chien, P.-C.; Huang, C.-Y.; Hsieh, H.-Y.; Hsieh, C.-H. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. *PLoS ONE* **2018**, *13*, e0207192. [CrossRef] [PubMed]
159. Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; Rahwan, I. The Moral Machine experiment. *Nature* **2018**, *563*, 59–64. [CrossRef] [PubMed]
160. Wang, J.; Ji, J.; Zhang, M.; Lin, J.-W.; Zhang, G.; Gong, W.; Cen, L.-P.; Lu, Y.; Huang, X.; Huang, D.; et al. Automated Explainable Multidimensional Deep Learning Platform of Retinal Images for Retinopathy of Prematurity Screening. *JAMA Netw. Open* **2021**, *4*, e218758. [CrossRef] [PubMed]
161. Zuallaert, J.; Godin, F.; Kim, M.; Soete, A.; Saeys, Y.; De Neve, W. SpliceRover: Interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics* **2018**, *34*, 4180–4188. [CrossRef]
162. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv* **2016**, arXiv:1606.05386.
163. Visani, G.; Bagli, E.; Chesani, F. OptiLIME: Optimized LIME explanations for diagnostic computer algorithms. *arXiv* **2020**, arXiv:2006.05714.
164. Salih, A.; Galazzo, I.B.; Raisi-Estabragh, Z.; Petersen, S.E.; Gkontra, P.; Lekadir, K.; Menegaz, G.; Radeva, P. A new scheme for the assessment of the robustness of Explainable Methods Applied to Brain Age estimation. In Proceedings of the 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Aveiro, Portugal, 7–9 June 2021.
165. Shickel, B.; Loftus, T.J.; Adhikari, L.; Ozrazgat-Baslanti, T.; Bihorac, A.; Rashidi, P. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. *Sci. Rep.* **2019**, *9*, 1879. [CrossRef]
166. Hartono, P. A transparent cancer classifier. *Health Inform. J.* **2018**, *26*, 190–204. [CrossRef] [PubMed]
167. Park, S.; Kim, Y.J.; Kim, J.W.; Park, J.J.; Ryu, B.; Ha, J.-W. Interpretable Prediction of Vascular Diseases from Electronic Health Records via Deep Attention Networks. In Proceedings of the IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 29–31 October 2018; pp. 110–117.
168. Bernardini, M.; Romeo, L.; Misericordia, P.; Frontoni, E. Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 235–246. [CrossRef] [PubMed]
169. Ming, Y.; Qu, H.; Bertini, E. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 342–352. [CrossRef] [PubMed]
170. Xiao, C.; Ma, T.; Dieng, A.B.; Blei, D.M.; Wang, F. Readmission prediction via deep contextual embedding of clinical concepts. *PLoS ONE* **2018**, *13*, e0195024. [CrossRef] [PubMed]
171. Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.* **2017**, *65*, 211–222. [CrossRef]
172. Silva, P.E.; Maldaner, V.; Vieira, L.; de Carvalho, K.L.; Gomes, H.; Melo, P.; Babault, N.; Cipriano, G.; Durigan, J.L.Q. Neuromuscular electrophysiological disorders and muscle atrophy in mechanically-ventilated traumatic brain injury patients: New insights from a prospective observational study. *J. Crit. Care* **2018**, *44*, 87–94. [CrossRef]
173. Menegotto, A.B.; Becker, C.D.L.; Cazella, S.C. Computer-aided diagnosis of hepatocellular carcinoma fusing imaging and structured health data. *Health Inf. Sci. Syst.* **2021**, *9*, 20. [CrossRef]

174. Simpson, S.; Kaufmann, M.C.; Glozman, V.; Chakrabarti, A. Disease X: Accelerating the development of medical countermeasures for the next pandemic. *Lancet Infect. Dis.* **2020**, *20*, e108–e115. [CrossRef]
175. Higgins, M.K. Can we AlphaFold our way out of the next pandemic? *J. Mol. Biol.* **2021**, *433*, 167093. [CrossRef] [PubMed]
176. Li, J.; Zhang, S.; Li, B.; Hu, Y.; Kang, X.-P.; Wu, X.-Y.; Huang, M.-T.; Li, Y.-C.; Zhao, Z.-P.; Qin, C.-F.; et al. Machine Learning Methods for Predicting Human-Adaptive Influenza A Viruses Based on Viral Nucleotide Compositions. *Mol. Biol. Evol.* **2019**, *37*, 1224–1236. [CrossRef]

Paper 2

electronics

MDPI

Article

Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log

Haider Ali [1,2], Imran Khan Niazi [3,4,5], David White [1,2], Malik Naveed Akhter [5] and Samaneh Madanian [1,*]

[1] Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand; haider.ali@aut.ac.nz (H.A.); david.white@aut.ac.nz (D.W.)
[2] Biodesign Lab, New Zealand College of Chiropractic, Auckland 1010, New Zealand
[3] Center of Chiropractic Research, New Zealand College of Chiropractic, Auckland 1010, New Zealand
[4] Center for Sensory-Motor Interaction, Department of Health Science and Technology, Aalborg University, 9220 Aalborg, Denmark
[5] Department of Clinical Sciences, Auckland University of Technology, Aukcland 1010, New Zealand; pjt6828@autuni.ac.nz
* Correspondence: sam.madanian@aut.ac.nz; Tel.: +64-(09)-921-9999 (ext. 6539)

Abstract: This study examines the performance of various machine learning (ML) models in predicting Interstitial Glucose (IG) levels using data from wrist-worn wearable sensors. The insights from these predictions can aid in understanding metabolic syndromes and disease states. A public dataset comprising information from the Empatica E4 smart watch, the Dexcom Continuous Glucose Monitor (CGM) measuring IG, and a food log was utilized. The raw data were processed into features, which were then used to train different ML models. This study evaluates the performance of decision tree (DT), support vector machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), lasso cross-validation (LassoCV), Ridge, Elastic Net, and XGBoost models. For classification, IG labels were categorized into high, standard, and low, and the performance of the ML models was assessed using accuracy (40–78%), precision (41–78%), recall (39–77%), F1-score (0.31–0.77), and receiver operating characteristic (ROC) curves. Regression models predicting IG values were evaluated based on R-squared values (−7.84–0.84), mean absolute error (5.54–60.84 mg/dL), root mean square error (9.04–68.07 mg/dL), and visual methods like residual and QQ plots. To assess whether the differences between models were statistically significant, the Friedman test was carried out and was interpreted using the Nemenyi post hoc test. Tree-based models, particularly RF and DT, demonstrated superior accuracy for classification tasks in comparison to other models. For regression, the RF model achieved the lowest RMSE of 9.04 mg/dL with an R-squared value of 0.84, while the GNB model performed the worst, with an RMSE of 68.07 mg/dL. A SHAP analysis identified time from midnight as the most significant predictor. Partial dependence plots revealed complex feature interactions in the RF model, contrasting with the simpler interactions captured by LDA.

Keywords: wearable sensors; machine learning; interstitial glucose; Empatica E4

Citation: Ali, H.; Niazi, I.K.; White, D.; Akhter, M.N.; Madanian, S. Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log. *Electronics* **2024**, *13*, 3192. https://doi.org/10.3390/electronics13163192

Academic Editor: Shing-Hong Liu

Received: 8 July 2024
Revised: 27 July 2024
Accepted: 8 August 2024
Published: 12 August 2024

1. Introduction

Diabetes is characterized by increased glucose levels. The incidence of it is increasing at a rapid rate. According to the World Health Organization, the number of people with diabetes worldwide rose from 108 million in 1980 to 422 million in 2014 [1], and this number is projected to reach 700 million by 2045 [2]. Prediabetes is a series of risk factors of diabetes defined using fasting glucose levels between 100 and 125 mg/dL [3]. Prediabetes affects approximately 34% of adults in the United States [3], with nearly 7.3 million undiagnosed cases [4,5]. However, 85% of individuals with prediabetes are unaware [6] that they have it [7]. Early intervention through lifestyle changes or medication can significantly reduce the risk of progression from prediabetes to diabetes by up to 58% [8]. Monitoring glucose

levels is thus helpful for managing and preventing metabolic diseases [9]. Classically, glucose levels are measured using a blood test that measures glycated hemoglobin levels (HbA1C). Fasting HbA1C levels measure glucose regulation for the past two to three months and do not measure fluctuations and short-term glucose spikes [10]. Monitoring short-term glucose variations is essential for adjusting medication, dietary habits, and physical activity to maintain optimal glucose regulation. To measure these short-term glucose spikes, continuous glucose markers (CGMs) are used. Glucose regulation markers such as time in range (TIR) can be measured using CGMs. CGMs are attached to the body with the help of a thread that penetrates the interstitial fluid. CGMs log Interstitial Glucose (IG) values every one to five minutes depending on the device. IG values are stored in them for up to 8 h. The stored IG values from CGMs can be downloaded with the help of Bluetooth technology [11]. CGMs require regular downloading of the data and are minimally invasive.

In comparison to CGMs, smart watches are noninvasive and self-updating. Therefore, there is an increased interest in using smart watches for predicting IG levels. An example of this growing interest is the curation of various datasets [12] that include smart watch data paired with glucose labels [13]. Smart watches are equipped with sensors capable of tracking various physiological parameters. Smart watch sensors include heart rate, an accelerometer, and skin conductance, etc. The smart watch sensor values can be used to engineer predictors of IG [14].

In addition to enhancing individual glucose management, predicting IG levels from smart watch sensors can also contribute to population-level health insights and disease management strategies. Aggregated data from smart watches and CGMs can provide valuable epidemiological information about glucose trends, the prevalence of metabolic conditions, and the impact of lifestyle factors on glycemic control.

Machine learning (ML) algorithms are used to predict IG markers from smart watches due to their ability to extract complex patterns and relationships [15]. ML models can adapt and improve over time by continuously learning from new data, making them well suited for personalized glucose monitoring and management [16]. Studies have demonstrated the effectiveness of ML algorithms in predicting glucose levels from wearable sensor data, achieving high levels of accuracy and precision [1,13,15,17]. Many of these models add food log data, in addition to smart watch data.

In this study, we categorize continuous glucose monitoring (CGM) values into high, low, and normal labels as described in [13,17,18]. Unlike traditional classifications of hyper- and hypoglycemia, which are tailored for diabetic patients, this approach considers individualized glucose fluctuations. These designations are dynamic and personalized, reflecting an individual's unique glycemic baseline and accounting for circadian and intra-/inter-day variability.

Related Work

Earlier works utilized support vector machines (SVMs) and decision trees for predicting IG values and categories using smart watch data. For example, ref. [18] produced 69 features predictive of glucose, defined the classification problem, and used decision trees to perform a classification with a root mean squared error (RMSE) equal to 21.22 ± 4.14 mg/dL. Similar works used depth vision guiding for recognizing human activity that can be used as input to glucose-monitoring models [19,20]. However, this requires additional sensors. Another work [13] recently designed four additional features using smart watch data but only performed a classification of CGM values into normal, high, and low and found support vector machine (SVM) to have an accuracy of 69% and decision tree (DT) to be 72.38% accurate. Another work utilizes extreme gradient boosting (XGBoost) models to classify the IG values of each participant with minimum accuracy = 60% and maximum accuracy = 86% [15]. While these works used smart watches and CGM data to train ML models to predict IG markers (classes and values), it would be useful to compare different ML models for both the classification and regression problem using the same

performance metrics. While these works report a hyperparameter tuning process for the models, there is a need for hyperparameter tuning for the best performing models. To compare the performance of the ML models, model explanations and visual techniques can inform why certain models such as tree models outperform other models [18]. The comparisons of earlier works is given in Table 1.

Table 1. The performance of different models in related works.

Study	Type	Models	Performance
[18]	Classification/Regression	DT	MSE = 21.22 ± 4.14 mg/dL
[13]	Classification	DT/SVM	Accuracy SVM (69%), DT (72.38%)
[15]	Classification	XGBoost	Accuracy (60–86%) *
[21]	Regression	Gradient Boosting	MSE = 23.40 mg/dL
[22]	Classification	DT	AUROC = 0.76 ± 0.07
[23]	Regression	RF	RMSE = 26.83 mg/dL
[14]	Classification	SVM	Accuracy = 72.6 ± 2.4%

* Individual specific models.

In summary, this work makes these novel contributions in comparison to other works:

- C1: Compare performance of ML models in predicting IG values using smart watch data as input and compared using Friedman test with Neymani post hoc analysis;
- C2: Compare performance of ML models in classifying IG values into high, low, and normal classes using smart watch data as input and compared using Friedman test with Neymani post hoc analysis;
- C3: Explain why different model types, such as tree-based models (RF, DT, and XGBoost), outperform SVM and GNB using partial dependence plots and Cook's plots.

In this context, our study compares several ML models, including DT, SVM, RF, Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), lasso cross-validation (LassoCV), Ridge, Elastic Net, and XGBoost. For the classification task, accuracy, precision, recall, F1-score, and ROC are used to compare models. Additionally, regression models predicting IG values are evaluated using R-squared values, mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and mean squared logarithmic error (MSLE). The hyperparameters of the best performing models are optimized using Bayesian Optimization with Optuna [24]. This work also explains why different models perform better than others, using partial dependence plots (PDP) to show feature interactions. This work also shows the robustness of RF models to influential outliers and the existence of such outliers using a Cook's plot.

The paper is organized as follows: Section 2 Materials and Methods: A detailed description of the dataset, highlighting its key characteristics and pertinent attributes, and an explanation of the data preprocessing, feature extraction techniques, and machine learning models used in this study. Section 3 Results: Presentation and comparison of the outcomes from the regression and classification analyses and their limitations. Sections 4 and 5 Discussion and Conclusion: Synthesis of the findings, discussion of their implications. The structure of paper is represented in Figure 1.

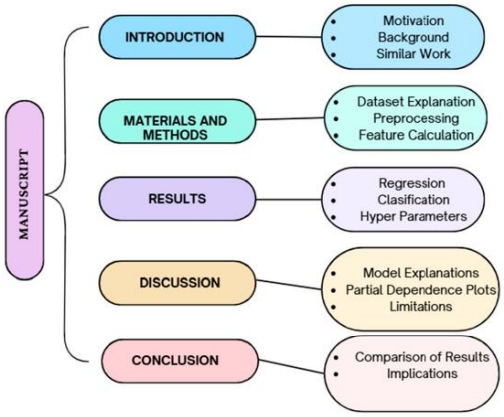


Figure 1. Structure of the manuscript.

2. Materials and Methods

In this study, we utilized a public dataset [12]. These data comprise a cohort of participants aged 35–65 years, inclusive, with elevated blood glucose levels falling within the range of normal to prediabetic. The dataset consists of 9 female participants and 7 male participants. The demographics are given in Supplementary Table S1. In this dataset, participants were required to wear a Dexcom G6 CGM and an Empatica E4 wristband for a duration of 8–10 days, during which physiological measurements such as heart rate, electrodermal activity, skin temperature, and tri-axial accelerometry were recorded.

Participants were also provided standardized breakfast meals every other day, and a food log was maintained. Date shifting was performed on the collected data to ensure participant de-identification.

The dataset includes a total of 16 participants. The files contain timestamped data. The ACC file provides accelerometer data for the X, Y, and Z orientations, while the BVP file records blood volume pulse measurements. The Dexcom, EDA, TEMP, IBI, and HR files contain data of IG values, electrodermal activity, skin temperature, inter-beat interval, and heart rate values, respectively. The food log file documents the food items consumed by each participant, including details such as date, time, logged food, amount, calories, total carbohydrates, dietary fiber, sugar, protein, and total fat content. Demographics, including gender and HbA1C values for each participant, are also provided. The PPG is sampled at 64 Hz, giving the HR and BVP every second for IBI computation. The EDA and skin temperature are sampled at 4 Hz, and the accelerometry at 32 Hz. CGM records a value of IG every 5 min.

For preprocessing, the HR and IBI data were filtered with a Chebyshev II order-4 filter with a stopband attenuation of 20 dB and a passband of 0.5–5 Hz, as described in [25]. For the removal of noise, a Gaussian low-pass filter was used with a sigma value of 400 ms [26]. We then segmented the filtered data into 5 min windows and aligned them with the Dexcom sensor values. Features were extracted from each window as described in [18]. These features can be broadly categorized into circadian features, statistical features of the sensor values, EDA features, and food features. A five-minute window was used for the calculation of these features, as the IG ground truth is available every five minutes. For the classification of IG labels, daily averages and standard deviations of CGM values were calculated. CGM values that are higher than the mean + standard deviation are considered high; conversely, values smaller than mean–standard deviation are considered low, whereas all the other values are considered normal, as described in [18].

Accelerometer data were preprocessed using a Butterworth low-pass filter with cutoff frequency = 20 Hz, as explained in [27]. The resultant acceleration was calculated from the X, Y, and Z components and corrected for the gravitational acceleration in the y direction. The EDA sensor data were smoothed to remove any artefacts. To do this, a Gaussian low-pass filter is used, with a 40-point window and value of sigma = 400 ms [26]. The HR data are filtered using a band-pass filter, filtering activity outside the [0.5–4 Hz] range. IBI data are filtered using a filter defined in [28]. BVP data are filtered using a moving average smoothing filter, whereas temperature sensor data are filtered using a Savitzky-Golay filter [29]. This is explained in Figure 2.

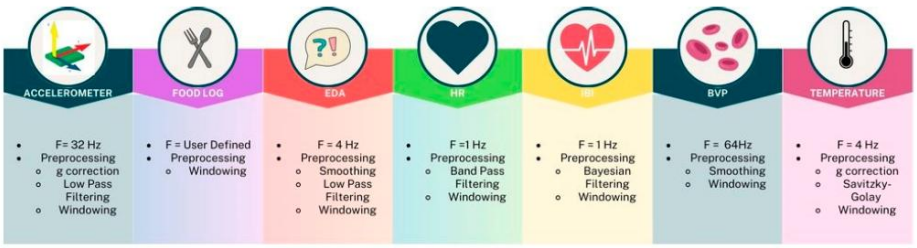


Figure 2. Preprocessing steps for each data source.

The features are calculated in this work as described in [13,18]. The features are broadly categorized into four main classes: food features, circadian features, statistical features, and autonomic nervous system features. These features are calculated for each 5 min window. The mathematical definition of the features is provided in Table 2.

Table 2. Features used in this work.

Feature	Description	Mathematical Expression
Biological Sex		
HbA1C	Glycated hemoglobin usually measured before the longitudinal data collection	
Mean of EDA, HR, IBI, T, and a	The mean of S (sensor value for the window for prediction usually equal to 5 min)	$\mu_S = \frac{\sum_{i=0}^{N} S}{N}$
Standard Deviation of EDA, HR, IBI, T, and a	The standard deviation of the S values for the length of the window	$\sigma_S = \sqrt{\frac{\left(\sum_{i=0}^{N} \mu_S - S\right)^2}{N}}$
Minimum Value of EDA, HR, IBI, T, and a		$\min_N S$
Maximum Value of EDA, HR, IBI, T, and a		$\max_N S$
First Quartile of EDA, HR, IBI, T and a	The value of 25% data point when the data are arranged in ascending order	$S(I)$ where I is the index of the S values in t ascending order rounded off to the nearest integer $I = \frac{N+1}{4}$
Third Quartile of EDA, HR, IBI, T and a	The value of 75% data point when the data are arranged in ascending order	$S(I)$ where I is the index of the S values in ascending order rounded off to the nearest integer $I = \frac{(N+1)x3}{4}$
Skewness of EDA, HR, IBI, T, and a	It is a measure of how symmetric the data are from the mean	$S = \frac{\left(\sum_{i=0}^{N} \mu_S - S\right)^3}{(N-1)\sigma_S^3}$
Peak of EDA values	Peak of EDA values in the prediction window for peaks of prominence 0.3	$\sum_{i=0}^{N} P$
Rolling mean of two hours of EDA values		
Rolling sum of 2 h of EDA peaks		
Standard Deviation of IBI (SDNN)	It is a measure of heart rate variability	
Root mean square of successive differences of inter-beat interval (RMSSD)	It is a measure of heart rate variability that is related to autonomic nervous system tone	$\mathrm{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N-1}[\mathrm{IBI}_{n+1} - \mathrm{IBI}_n]^2}$

Table 2. *Cont.*

Feature	Description	Mathematical Expression
Number of times IBI exceeds 50 ms (NN50)	It is a measure of heart rate variability and sympathetic nervous system activation	$NN_{50} = \sum_{i=1}^{n} a$ where a = 1 if $IBI_{i+1} - IBI_i > 50$ Else a = 0
pNN50	It is the measure of how many times in time N the IBI has exceeded 50 ms	$pNN_{50} = \frac{NN_{50}}{N}$
Calorie Sums	Rolling sums of 2 h, 8 h, and 24 h of calorie estimates are used	
Protein Sums	Rolling sums of 2 h, 8 h, and 24 h of protein consumption estimates are used	
Carbohydrate Sums	Rolling sums of 2 h, 8 h, and 24 h of carbohydrate consumption estimates are used	
Sugar Sums	Rolling sums of 2 h, 8 h, and 24 h of sugar consumption estimates are used	
Rolling mean of two-hour acceleration values	Used to estimate activity levels	
Rolling maximum value of two-hour acceleration values	Used to estimate activity levels	
Activity Bouts	When the mean of the window exceeds the rolling mean of acceleration values	
Individual Number	Used to model individuality	

The efficacy of the features was verified based on correlation and mutual information for discrete and t-Distributed Stochastic Neighbor Embedding (t-SNE) plots. Correlation is used to measure the independent features. These features are used to train ML models.

DT is a non-parametric ML algorithm used for classification and regression tasks. It splits the data into subsets based on the value of the input features, forming a tree structure where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome. The goal is to create a model that predicts the target variable by learning simple decision rules inferred from the data features.

SVM finds the hyperplane that best separates the classes in the feature space, maximizing the margin between the closest points of the classes (support vectors). SVM is effective in high-dimensional spaces and is particularly useful for problems where the number of dimensions exceeds the number of samples.

RF is an ensemble learning method that combines multiple decision trees to improve predictive performance and control overfitting. Each tree is trained on a random subset of the data and the features. The final prediction is made by averaging the outputs of individual trees (for regression) or by majority voting (for classification).

LDA is a dimensionality reduction technique used for classification. It projects the data onto a lower-dimensional space where the classes are most separable. LDA assumes that the features follow a Gaussian distribution and that different classes have identical covariances. It finds linear combinations of the features to predict the labels.

KNN is a simple, instance-based learning algorithm used for classification and regression. It predicts the target by finding the K training samples closest in distance to a new data point and returning the majority class (for classification) or the average value (for regression). KNN is non-parametric and makes predictions based on the entire dataset.

GNB is a probabilistic classifier based on Bayes' theorem, assuming independence between features given the class. It models the distribution of the features within each class as Gaussian.

LassoCV is a linear regression model that includes L1 regularization (lasso) to enforce sparsity, reducing the number of features by shrinking some coefficients to zero. The CV part stands for cross-validation, which is used to find the optimal regularization parameter. It helps prevent overfitting by penalizing the absolute size of the coefficients.

Ridge regression is a linear regression model with L2 regularization, which penalizes the squared magnitude of the coefficients. This regularization helps to prevent overfitting by shrinking the coefficients towards zero but unlike lasso, it does not enforce sparsity. It is useful when dealing with multicollinearity or when the number of predictors exceeds the number of observations.

AdaBoost, short for Adaptive Boosting, is an ensemble learning technique that combines multiple weak classifiers to create a strong classifier. It works by iteratively training classifiers on weighted versions of the data, where the weights are adjusted to focus on the hardest-to-classify samples. The final model is a weighted sum of the individual classifiers.

XGBoost (extreme gradient boosting) is an efficient and scalable implementation of the gradient boosting framework. It builds an ensemble of decision trees in a sequential manner, where each tree corrects the errors of its predecessor. XGBoost uses advanced regularization techniques to reduce overfitting and includes features like parallel tree construction, handling missing values, and optimized computations.

For both regression and classification, numerical features were normalized using the Z-score. The Z-score is defined in Equation (1).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x gives the value of the feature, μ is the mean of the feature's distribution, and σ is its standard deviation. There are two categorical features in this dataset: participant ID and gender. One-hot encoding is a method used to transform categorical data into a numerical format suitable for ML models. It converts each category into a unique binary vector, where only one element is set to 1 and the rest are 0. In this work, the categorical features are one-hot encoded. The ML models are trained using a subset of data called training data, learning to predict IG values with respect to the input smart watch and food log features. Features from an independent subset called validation data are used to predict IG values; these values are compared with actual IG values for determining the performance of the model.

The following performance metrics are used in the comparison of the regression models.

1. MAE: MAE measures the average absolute difference between the predicted and actual IG values. It is less sensitive to outliers;
2. RMSE: It is the root of the average squared difference between the predicted and actual IG values;
3. MAPE: MAPE measures the average ratio of error (difference between the actual and predicted value) with the actual IG value;
4. R-squared (R^2) and Adjusted R-squared: R^2 measures the proportion of variance in the CGM values explained by the input smart watch features. A higher R^2 indicates a better fit of the model to the data;
5. MSLE: It is the average difference between the log of the actual and predicted IG values. It is specifically less sensitive to outliers.

For classification, the following parameters are used to define the performance of the models. A true positive (TP) occurs when the model correctly predicts a positive class for an instance that is actually positive. A true negative (TN) is when the model correctly predicts a negative class for an instance that is actually negative. A false positive (FP), also known as a Type I error, happens when the model incorrectly predicts a positive class for an instance that is actually negative. Conversely, a false negative (FN), or Type II error, occurs when the model incorrectly predicts a negative class for an instance that is actually positive.

The following performance metrics are used to compare the classification models.

1. Accuracy (%): Accuracy measures the proportion of correctly classified instances out of the total instances. It is given by (TP+TN/TP+TN+FP+FN);
2. Precision: Precision, also known as positive predictive value, measures the proportion of true positive predictions among all positive predictions made by the model. It is given as (TP/TP+FP);
3. Recall: Recall, also known as sensitivity or the true positive rate, measures the proportion of true positives identified by the model out of all actual positives. It is given as (TP/TP+FN);
4. F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall;

5. ROC (Receiver Operating Characteristic) Curve: The ROC curve plots the true positive rate (recall) against the false positive rate at various threshold settings. The ROC AUC (Area Under the Curve) measures the model's ability to discriminate between classes, with a higher AUC indicating a better performance.

To make sure that the difference between performance metrics is statistically significant, they are compared using the Friedman test. This is done by dividing the dataset into 10 folds, training it on 9 folds and testing on the remaining fold (once for each fold). The performance metrics are recorded, and these differences are interpreted using the Nemenyi post hoc test.

3. Results

3.1. Feature Calculation

These features are calculated in python using NumPy, pandas, and JAX.

The correlation heatmap in Figure 3 illustrates the relationships between various calculated features from the food log and smart watch data. Notably, some clusters of features, such as those related to proteins and carbohydrates over different time windows, exhibit strong correlations within their groups. Conversely, features like heart rate variability (HRV) metrics and activity measures demonstrate more independence, as indicated by their lighter shades.

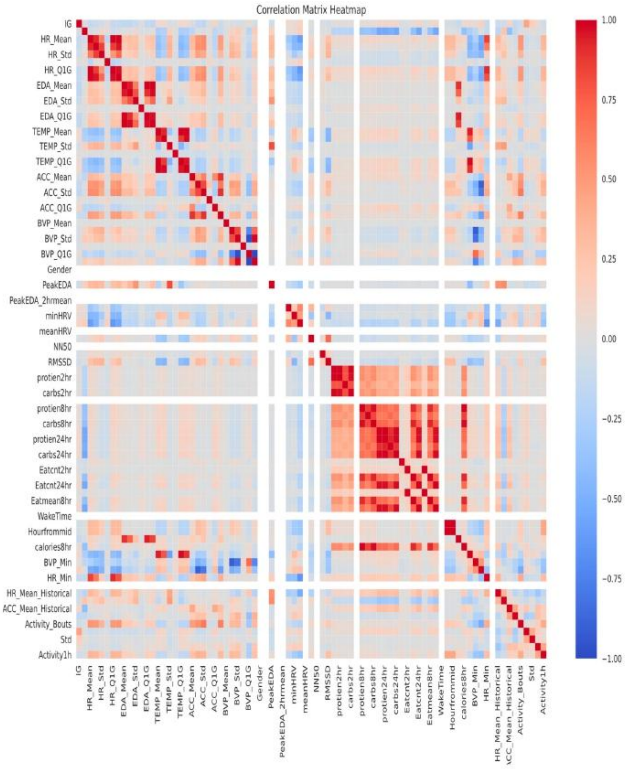


Figure 3. Correlation heatmap for all the calculated features. The stronger shades of red signify a positive correlation, and blue signifies a negative correlation. The lighter shades signify the features that have a smaller correlation, meaning that they are potentially independent.

Figure 4 illustrates the Pearson correlation coefficients between various features and IG values. Each feature's correlation with the dependent variable is visually represented, where positive correlations extend to the right and negative correlations to the left. Features such as activity counts, historical accelerometer data (ACC), heart rate (HR) metrics, and carbohydrate intake exhibit high degrees of correlation with the CGM values.

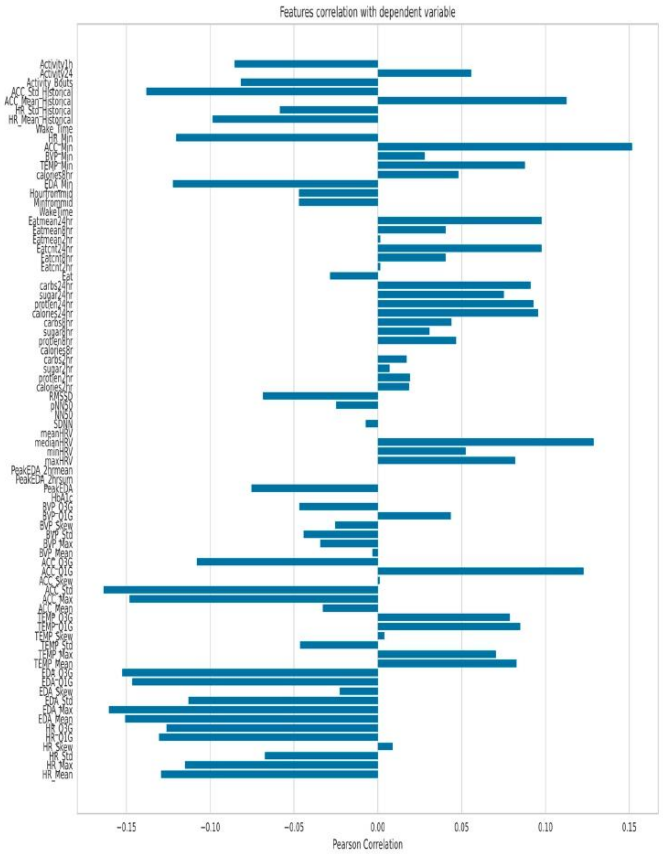


Figure 4. Feature correlation with Interstitial Glucose levels.

3.2. Regression

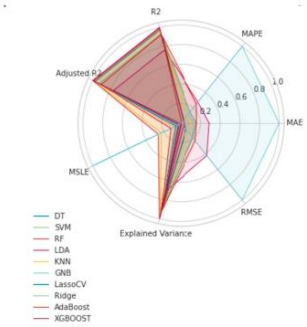
Regression models predict IG values based on smart watch and food log features. These ML models are used for IG prediction: DT, SVM, RF, LDA, KNN, GNB, LassoCV, Ridge, AdaBoost, and XGBoost. After identifying the best performing model type (RL), its hyperparameters were tuned using Bayesian Optimization.

The feature data were split into (70%) training and (30%) testing. The models were trained on the training set. The trained models were used to predict IG values on the unseen testing data. The predicted IG values were subtracted from the actual IG value to determine the *error* value. The error values were used to calculate various performance metrics of the regression models. Table 3 compares the performance metrics of the ML models in predicting IG values based on the input features.

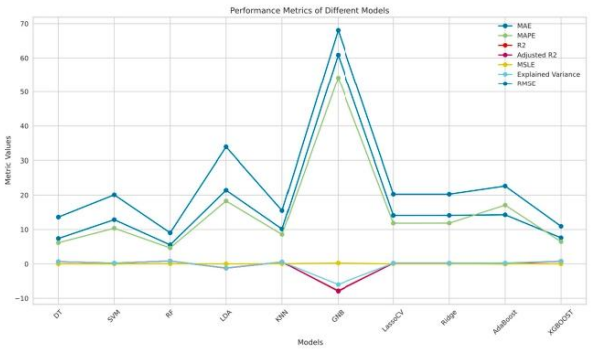
Table 3. Performance metrics of regression models.

Model	MAE (mg/dL)	MAPE	R^2	Adjusted R^2	MSLE (mg/dL)	Explained Variance	RMSE (mg/dL)
DT	7.379	6.15	0.64	0.64	0.0115	0.6475	13.61
SVM	12.86	10.37	0.23	0.21	0.004	0.25	20.09
RF	5.54	4.65	0.84	0.84	0.068	0.84	9.04
LDA	21.42	18.30	−1.19	−1.22	0.014	−1.16	33.94
KNN	10.14	8.57	0.54	0.53	0.06	0.54	15.51
GNB	60.85	54.18	−7.81	−7.96	0.25	−6.00	68.07
LassoCV	14.11	11.84	0.21	0.20	0.02	0.21	20.27
Ridge	14.12	11.85	0.21	0.20	0.025	0.21	20.27
AdaBoost	14.28	17.10	0.194	0.007	0.034	0.27	22.64
XGBoost	7.59	6.45	0.77	0.768	0.007	0.77	10.93

Figure 5 presents the performance parameters of the models visually. Residuals and QQ plots are plotted in the Supplementary Material. RF consistently has high performance across all metrics. The RF model has a high R^2, adjusted R^2, and explained variance. while maintaining a low MAE, MAPE, MSLE, and RMSE.



(**a**)



(**b**)

Figure 5. Comparison of the performance metrics of regression models: (**a**) Normalized spider plot for difference performance metrics of regression results; (**b**) bar plot for performance measures of different models.

To make sure that the difference between the metrics for each model is significant, a Friedman test is carried out for each metric. The results for each metric and model are reported in the Supplementary Material. An example of this analysis is shown here for reference. The Friedman statistic for values of MAE for all the models for all folds is 70.0, with ($p = 1.47 \times 10^{-12}$) showing that the difference is significant. To understand for which models the comparison is significant, a Nemenyi post hoc analysis for the results is conducted and plotted as a heatmap in Figure 6.

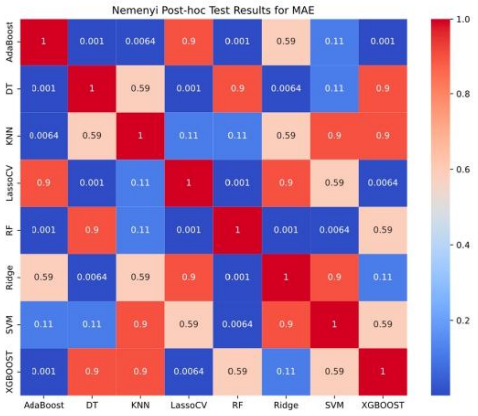
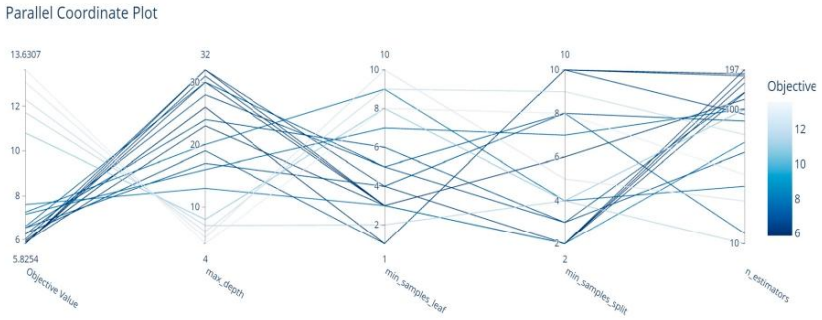


Figure 6. Nemenyi post hoc analysis of the Friedman test for MAE across all the models.

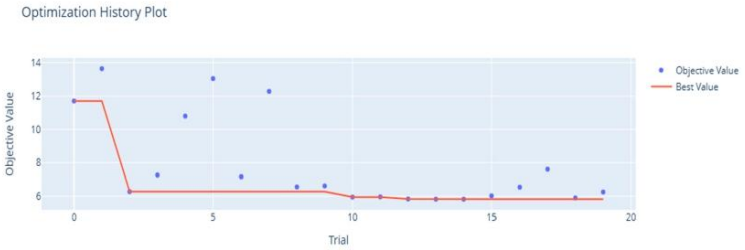
Based on the Nemenyi post hoc test results, we can observe how different models compare against each other in terms of performance. The heatmap visualizes the pairwise comparisons of the models, highlighting the statistically significant differences in their MAE values. Each cell in the heatmap represents a comparison between two models. A smaller value signifies a more substantial difference in performance between the compared models.

The Nemenyi post hoc test results for mean absolute error (MAE) reveal significant and non-significant differences in model performance. Significant differences ($p < 0.05$) were observed between AdaBoost and decision tree ($p = 0.001$), AdaBoost and KNN ($p = 0.006$), AdaBoost and Random Forest ($p = 0.001$), AdaBoost and XGBoost ($p = 0.001$), decision tree and LassoCV ($p = 0.001$), decision tree and Ridge ($p = 0.006$), Random Forest and LassoCV ($p = 0.001$), Random Forest and Ridge ($p = 0.001$), SVM and Random Forest ($p = 0.006$), and XGBoost and LassoCV ($p = 0.006$). No significant differences ($p \geq 0.05$) were found between AdaBoost and LassoCV, AdaBoost and Ridge, AdaBoost and SVM, decision tree and KNN, decision tree and Random Forest, decision tree and SVM, decision tree and XGBoost, KNN and LassoCV, KNN and Random Forest, KNN and Ridge, KNN and SVM, KNN and XGBoost, LassoCV and Ridge, LassoCV and SVM, Random Forest and XGBoost, Ridge and SVM, Ridge and XGBoost, and SVM and XGBoost.

The parameters of the RF model are tuned using Optuna. Optimal hyperparameters are reported in Table 2. Figure 7 shows the optimization process. The number of estimators is the number of decision trees in the RF model, maximum depth is the maximum depth of each decision tree, and minimum sample leaf is the smallest number of samples that should be present in the leaf node after splitting a node. This hyperparameters are reported in Table 4.



(**a**)



(**b**)

Figure 7. Bayesian Optimization for hyperparameter tuning: (**a**) Parallel coordinates shaded with the objective value; the objective for the optimization is the RMSE value. (**b**) The evolution of the RMSE over the number of iterations.

Table 4. Optimal hyperparameters measured using Bayesian Optimization using Optuna.

Hyperparameter	Number of Estimators	Maximum Depth	Minimum Sample Split	Minimum Leaves Per Sample
Value	178	26	10	1

3.3. Classification

The classification models categorize IG values into normal, high, and low. Most IG values belong to the normal class in this dataset. To overcome this class imbalance, the number of samples per class is stratified by downsampling the normal class to 2500 samples. The total number of samples is 7500 (2500 samples per class). A total of 70% of the samples are used in training the ML models, while 30% are used for testing the performance of those models (RF, DT, SVM, LDA, KNN, GNB, Ridge, AdaBoost and XGBoost). After identifying the best performing model type (RL), its hyperparameters are tuned using Bayesian Optimization.

Figure 8 represents the performance of classification models in terms of accuracy, precision, recall and F1-score. Table 5 presents the comparison of different classification models in glucose prediction.

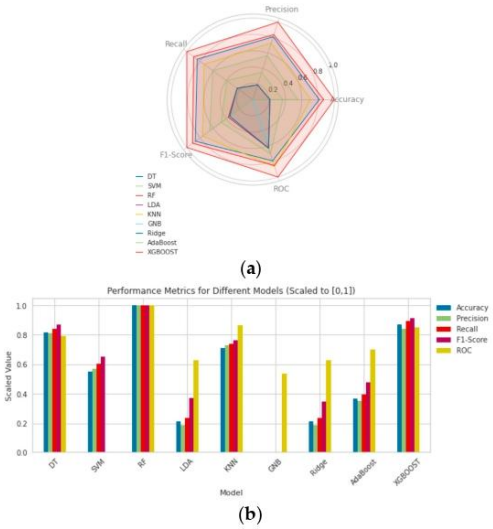


Figure 8. Comparison of the performance metrics of classification models: (**a**) Normalized spider plot for different performance metrics of classification; (**b**) bar plot for performance measures of different models.

Table 5. Performance metrics of classification models.

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC
DT	71	71	71	0.71	0.78
SVM	61	62	62	0.61	0.25
RF	78	78	77	0.77	0.92
LDA	48	48	48	0.48	0.67
KNN	67	68	67	0.66	0.83
GNB	40	41	39	0.31	0.61
Ridge	48	48	48	0.47	0.67
AdaBoost	54	54	54	0.53	0.72
XGBoost	73	72	73	0.73	0.82

To make sure that the difference between the metrics for each model is significant, a Friedman test is carried out for each metric. The results for each metric and model are reported in the Supplementary Material. An example of this analysis is shown here for reference. The Friedman statistic for values of accuracy for all the models for all folds is 78.33 with ($p = 1.05 \times 10^{-13}$), showing that the difference is significant. To understand for which models the comparison is significant, a Nemenyi post hoc (Figure 9) analysis for the results is conducted and plotted as a heatmap.

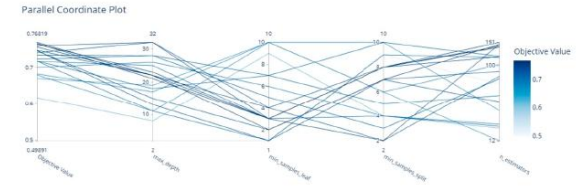
The Nemenyi post hoc test results for mean absolute error (MAE) reveal significant and non-significant differences in model performance. Significant differences ($p < 0.05$) were observed between AdaBoost and XGBoost ($p = 0.001$), decision tree and GNB ($p = 0.001$), GNB and KNN ($p = 0.006$), Random Forest and GNB ($p = 0.001$), GNB and XGBoost ($p = 0.001$), LDA and RF ($p = 0.002$), and LDA and XGBoost ($p = 0.001$). No significant differences ($p \geq 0.05$) were found between AdaBoost and DT, AdaBoost and GNB, AdaBoost and KNN, AdaBoost and LDA, decision tree and KNN, decision tree and Random Forest, decision tree and SVM, decision tree and XGBoost, KNN and LassoCV, KNN and Random Forest, KNN and Ridge, KNN and SVM, KNN and XGBoost, LassoCV and Ridge, LassoCV and SVM, Random Forest and XGBoost, Ridge and SVM, and SVM and XGBoost.

Figure 9. Nemenyi post hoc test results for accuracy (%).

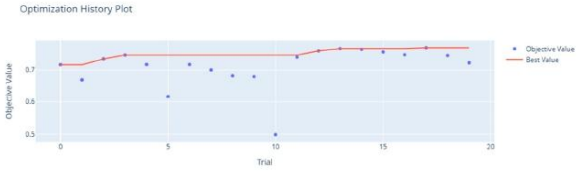
RF outperformed the other models across all the performance metrics. Other performance plots are provided in the Supplementary Material for further clarification. The class prediction error, confusion matrix, ROC curves, and precision recall curves of the classifier were trained using the tuned hyperparameters given in Table 6 in Figure 10.

Table 6. Optimal hyperparameters measured using Bayesian Optimization using Optuna.

Hyperparameter	Number of Estimators	Maximum Depth	Minimum Sample Split	Minimum Leaves Per Sample
Value	130	22	7	2



(**a**)



(**b**)

Figure 10. Bayesian Optimization for hyperparameter tuning: (**a**) Parallel coordinates shaded with the objective value; the objective for the optimization is accuracy. (**b**) The evolution of the accuracy over the number of iterations.

Figure 11 represents the performance of an RF model trained on 70 % data and tested on 30% testing data.

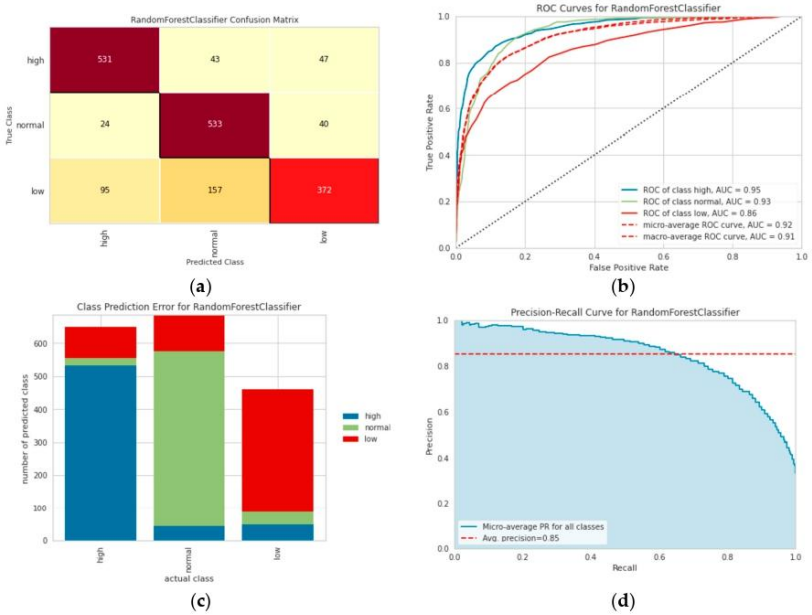


Figure 11. Performance of the tuned Random Forest model on validation data of the balanced dataset: (**a**) Confusion matrix of the tuned RF classifier for validation data of the balanced dataset, (**b**) ROC curves of the tuned RF classifier for validation data of the balanced dataset, (**c**) class prediction error of the tuned RF classifier for validation data of the balanced dataset, and (**d**) precision recall curve of the tuned RF classifier for validation data of the balanced dataset.

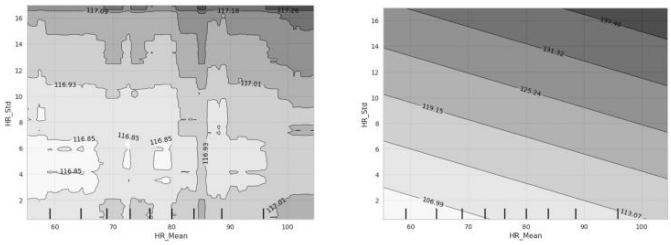
3.4. Model Explanations

The best regression and classification models are explained. To better understand why tree models perform better than kernel-based models such as SVMs, generative models such as GNB, and non-parametric models such as KNNs, partial dependence plots are used to show the complex interaction of the features modeled.

According to the literature, tree-based models are robust to noise and suitable for visualizing feature interactions. Here, we plot the partial dependence plots of two features we believe interact with each other in a complex manner (HR_Mean and HR_Std). As can be seen from the partial dependence plot (PDP) from the RF and the LDA in Figure 12, the PDP of the RF model represents a complex relationship, whereas for the linear model, the PDP shows a linear relationship, resulting in lower performance metrics for regression.

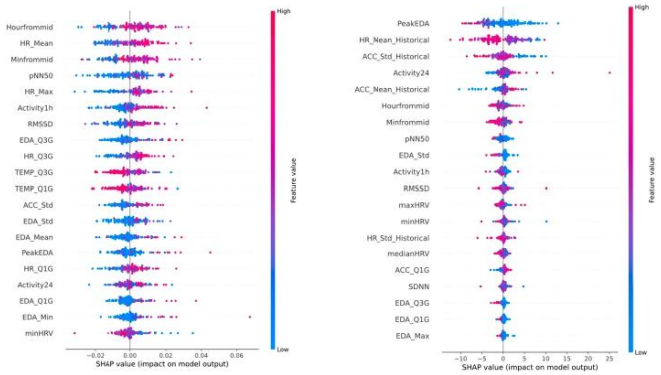
The SHAP plot in Figure 13a reveals the distribution and impact of features on the RF model's predictions for classifying CGM values into high, low, and normal categories. Each feature's influence is illustrated by the spread of dots along the *x*-axis, indicating the SHAP values. A wider spread of dots signifies a greater variance in the feature's impact across different data points. For instance, Hourfrommid and HR_Mean have a broad range of SHAP values, showing they can significantly sway the predictions towards both high and low CGM classes. The color gradient of the dots, from blue (low feature value) to red (high feature value), further elucidates how different levels of a feature affect the

prediction. This spread and color coding collectively depict the nuanced contributions of each feature, demonstrating the complex interplay between physiological metrics, food intake, and glucose levels. This shows that for the classification problem, the circadian features are relatively more important than other features. This is consistent with earlier works [13,15,18].



(**a**) (**b**)

Figure 12. Comparison of PDP plots for standard deviations of heart rate and mean heart rate: (**a**) The RF PDP captures a complex relationship, resulting in a higher accuracy; (**b**) the LDA assumes a linear relationship, resulting in a lower performance.



(**a**) (**b**)

Figure 13. SHAP summary plots for classification and regression. (**a**) SHAP values for classification, (**b**) SHAP values for regression.

Figure 11b demonstrates how each feature contributes to the regression model's prediction of CGM values, highlighting the importance and influence of physiological metrics and activity levels captured by the Empatica E4 smart watch, as well as historical data. Key features like '*PeakEDA*', '*HR_Mean_Historical*', and '*ACC_Std_Historical*' exhibit a broad range of SHAP values, indicating significant variability in their impact on CGM predictions. For instance, high values of '*PeakEDA*' tend to push predictions higher, while low values often push predictions lower. Similarly, '*HR_Mean_Historical*' consistently shows a positive impact on CGM values, with high feature values leading to higher predictions. The varied influence of '*ACC_Std_Historical*' and 'Activity24' further underscores the complexity and interplay of factors affecting glucose levels. This comprehensive visualization of feature contributions provides valuable insights into the model's decision-making process.

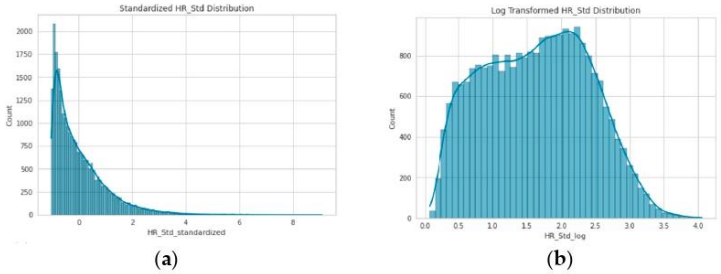
4. Discussion

For both classification and regression tasks RF, has a superior performance, while the other tree-based model, DT, is not that far behind. XGBoost, which is also a tree model, performs well in both the tasks as well. Tree models are known to perform well in cases when there are nonlinear relationships. KNNs [30] and tree-based models (RF, DT, and XGBoost) are both equipped to handle nonlinear relationships between the data [31–33].

Gaussian Naïve Bayes (GNB) is best suited for datasets with conditionally independent features, linear relationships, and normalized data [34]. While it assumes conditional feature independence, which simplifies computation, this assumption can limit its performance with more complex, dependent features. The feature interactions that GNB can model are also linear, which is not the case for the complex relationship between many of these variables—for example, the interaction between the rolling sum of carbohydrates consumed and hours from midnight. Food can be consumed at the beginning of the day, which is less far from midnight, but that potentially increases CGM values in the subsequent windows of prediction.

KNN relies on distance measurements, which can be less effective with mixed data types and skewed distributions. Since most of these variables are skewed, KNN underperforms the tree models but outperforms GNB.

Support vector machines (SVMs) can capture complex feature relationships but often require extensive feature engineering to perform optimally [35]. For instance, standardization and transformations such as taking the logarithm of skewed features can improve SVM performance [35]. Figure 14 illustrates the impact of standardization on the skewness of the heart rate (HR) standard deviation: normalization using the Z-score does not eliminate skewness, but applying a log transformation makes the changes more prominent. To illustrate the impact of skewness on model performance, models sensitive to skewness (SVM and GNB) are trained on the training set. The accuracy calculations on the validation set reveals that taking a log of HR_Std increases the accuracy of GNB (40% to 44%) and SVM (61% to 64%), although that increase is small.



(**a**) (**b**)

Figure 14. Comparison of HR standard deviation skewness. (**a**) Normalization of HR values using the Z-score does not eliminate the skewness of the data. (**b**) Taking a log of this value makes the changes more prominent.

Tree-based models are particularly adept at handling skewed features and mixed data types [36]. Of the tree models, ensemble methods like RF and XGBoost outperform other models by better handling outliers. The presence of influential outliers in the data is evidenced by the Cook's distance plot in Figure 15. RF's better performance over XGBoost in this study may be due to the noise in the wearable data, suggesting that quality metrics should be used during data collection to minimize noise influence.

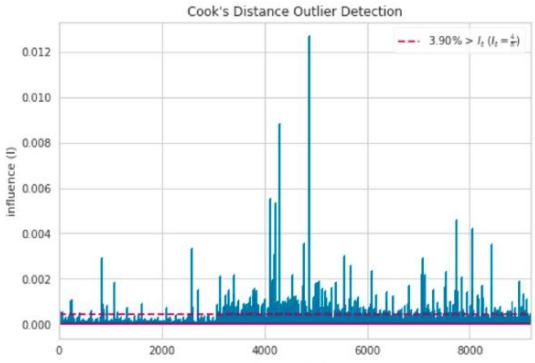


Figure 15. Cook's distance plot shows influential outliers.

Similarly, tree-based models are better at learning mixed data types (categorical and numerical) than KNN models, because the latter rely on distance measurements. Amongst the three models, ensemble methods (RF and XGBOOST) performed better, because of their better handling of outliers.

The preprocessing, feature engineering, and hyperparameter tuning in this work results in models with superior performance. Table 7 summarizes the comparison of the best performing models in this work whereas Table 8 presents the comparison of trained models with earlier works.

Table 7. Performance of the RF model trained on tuned hyperparameters.

Class	Precision	Recall	F1-Score	Accuracy
High	0.80	0.85	0.82	0.80
Normal	0.71	0.89	0.79	0.70
Low	0.81	0.58	0.67	0.80

Table 8. Comparison of trained models with similar work.

Study	Type	Model	Performance
[18]	Regression	DT	MSE = 21.22 $\pm$ 4.14 mg/dL
This Work	Regression	RF	MSE = 9.04 mg/dL
[14]	Classification	DT	AUROC = 0.72
This work	Classification	RF	AUROC = 0.86

Tree-based models are effective in predicting glucose levels both for classification and regression. The explanations also provide a basis for future model development and feature engineering; for example, it is worthwhile to convert the skewed features to the log of these features or using PDP plots to engineer new features.

This study has potential limitations. The values of ground truth measured using a Dexcom sensor that define the labels are affected by motion [37]. However, the study clearly states the values it predicts. The models that are compared in this study are compared based on the features that have been engineered. Future feature engineering or postprocessing of the features, such as taking a log of the features, can affect the performance of different model types.

These results underscore the importance of using robust ensemble methods for glucose level prediction, suggesting that these models can significantly improve the accuracy and reliability of real-time glucose-monitoring systems. In practical terms, the enhanced performance of these models can lead to better glucose management and improved health

outcomes for individuals with diabetes. Additionally, this study's insights into feature engineering, such as the benefits of log transformation for skewed data, provide a valuable framework for developing more accurate predictive models in future research. Implementing these findings in healthcare settings could facilitate more personalized and effective diabetes management, ultimately contributing to better patient care and quality of life.

5. Conclusions

This study has demonstrated that tree-based models, particularly Random Forest (RF) and decision tree (DT), exhibit superior performance in predicting Interstitial Glucose (IG) levels from wrist-worn wearable sensor data. These models outperformed other machine learning (ML) models in both classification and regression tasks, achieving higher accuracy, precision, recall, and F1-scores for classification, as well as lower root mean square error (RMSE) and higher R-squared values for regression.

In conclusion, the findings of this study highlight the potential of using wearable sensor data and tree-based ML models to provide insights into metabolic health and disease states. Future work should focus on improving data quality through noise reduction techniques and exploring advanced feature engineering methods, such as partial dependence plots (PDP) and transforming skewed features. Implementing these improvements can further enhance the accuracy and reliability of IG level predictions, ultimately contributing to the better management of metabolic syndromes and diseases.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics13163192/s1 Table S1: Demographics of the participants in the dataset, Table S2: Features Definition, Figure S1: Performance comparison of the regression models, Figure S2: QQ plot of the Regression models, Figure S3: Prediction Error of the Regression Models, Figure S4: Regression Model performance (kfold k = 10), Figure S5: t-SNE plot of the features, Figure S6: Confusion Matrix of the Classification Models, Figure S7: Classification Model performance (kfold k = 10), Figure S8: Nemenyi Post-hoc results for MAE, Figure S9: Nemenyi Post-hoc results for Adjusted R2, Figure S10: Nemenyi Post-hoc results for Explained Variance, Figure S11: Nemenyi Post-hoc results for MSLE, Figure S12: Nemenyi Post-hoc results for MAE, Figure S13: Nemenyi Post-hoc results for Accuracy, Figure S14: Nemenyi Post-hoc results for precision, Figure S15: Nemenyi Post-hoc results for F-1 Score, Figure S16: Nemenyi Post-hoc results for Recall, Figure S17: Nemenyi Post-hoc results for ROC-AUC.

Author Contributions: Conceptualization, H.A., I.K.N., S.M., M.N.A. and D.W.; methodology, I.K.N., M.N.A. and S.M.; software, H.A., I.K.N., M.N.A. and S.M.; validation, I.K.N., D.W. and S.M.; formal analysis, H.A., I.K.N., M.N.A. and S.M.; investigation, H.A.; resources, I.K.N. and D.W.; data curation, H.A.; writing—original draft preparation, H.A.; writing—review and editing, H.A., M.N.A., S.M. and D.W.; visualization, H.A.; supervision, I.K.N., S.M. and D.W.; project administration, D.W.; funding acquisition, I.K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data can be made available on reasonable request. The dataset is available at https://physionet.org/content/big-ideas-glycemic-wearable/1.1.1/ (accessed on 6 June 2024).

Acknowledgments: The authors acknowledge support by the New Zealand College of Chiropractic PhD Scholarship (funding number: 20126384), and computational support is provided by New Zealand E Science Infrastructure (NESI grant number AUT 03802).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ML	Machine Learning
IG	Interstitial Glucose
CGM	Continuous Glucose Monitoring
ACC	Accelerometer
BVP	Blood Volume Pulse
EDA	Electrodermal Activity
HR	Heart Rate
IBI	Inter-Beat Interval
DT	Decision Tree
SVR	Support Vector Regression
RF	Random Forest
LDA	Linear Discriminant Analysis
KNN	K-Nearest Neighbors
GNB	Gaussian Naïve Bayes
LassoCV	Lasso Cross-Validation
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error
MSLE	Mean Square Logarithmic Error
ROC	Receiver Operator Curve
AUROC (Area Under Receiver Operator Curve)	Area Under Receiver Operator Curve
PDP	Partial Dependence Plot
SHAP	Shapley Additive Explanations

References

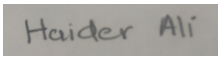
1. Maged, Y.; Atia, A. The Prediction Of Blood Glucose Level By Using The ECG Sensor of Smartwatches. In Proceedings of the 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 8–9 May 2022; pp. 406–411. [CrossRef]
2. Bent, B.; Cho, P.J.; Wittmann, A.; Thacker, C.; Muppidi, S.; Snyder, M.; Crowley, M.J.; Feinglos, M.; Dunn, J.P. Non-invasive wearables for remote monitoring of HbA1c and glucose variability: Proof of concept. *BMJ Open Diabetes Res. Care* **2021**, *9*, e002027. [CrossRef]
3. International Diabetes Federation. IDF Diabetes Atlas Tenth Edition 2021. Available online: https://diabetesatlas.org/ (accessed on 3 June 2024).
4. Aguilar, M.; Bhuket, T.; Torres, S.; Liu, B.; Wong, R.J. Prevalence of the Metabolic Syndrome in the United States, 2003–2012. *JAMA* **2015**, *313*, 1973–1974. [CrossRef] [PubMed]
5. CDC. National Diabetes Statistics Report, Diabetes. Available online: https://www.cdc.gov/diabetes/php/data-research/index.html (accessed on 3 June 2024).
6. Grundy, S.M.; Brewer, H.B., Jr.; Cleeman, J.I.; Smith, S.C., Jr.; Lenfant, C. Definition of Metabolic Syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation* **2004**, *109*, 433–438. [CrossRef] [PubMed]
7. Ervin, R.B. Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003–2006. *Natl. Health Stat. Rep.* **2009**, *13*, 1–7.
8. Ford, E.S.; Li, C.; Sattar, N. Metabolic syndrome and incident diabetes: Current state of the evidence. *Diabetes Care* **2008**, *31*, 1898–1904.
9. Jarvis, P.R.; Cardin, J.L.; Nisevich-Bede, P.M.; McCarter, J.P. Continuous glucose monitoring in a healthy population: Understanding the post-prandial glycemic response in individuals without diabetes mellitus. *Metabolism* **2023**, *146*, 155640. [CrossRef]
10. CDC. Prediabetes—Your Chance to Prevent Type 2 Diabetes, Diabetes. Available online: https://www.cdc.gov/diabetes/prevention-type-2/prediabetes-prevent-type-2.html (accessed on 3 June 2024).
11. Zoungas, S.; Chalmers, J.; Ninomiya, T.; Li, Q.; Cooper, M.E.; Colagiuri, S.; Fulcher, G.; De Galan, B.E.; Harrap, S.; Hamet, P.; et al. Association of HbA1c levels with vascular complications and death in patients with type 2 diabetes: Evidence of glycaemic thresholds. *Diabetologia* **2012**, *55*, 636–643. [CrossRef]
12. Beck, R.W.; Bergenstal, R.M.; Riddlesworth, T.D.; Kollman, C.; Li, Z.; Brown, A.S.; Close, K.L. Validation of Time in Range as an Outcome Measure for Diabetes Clinical Trials. *Diabetes Care* **2018**, *42*, 400–405. [CrossRef]
13. Cho, P.; Kim, J.; Bent, B.; Dunn, J. BIG IDEAs Lab Glycemic Variability and Wearable Device Data. *PhysioNet* **2023**.

14. Ali, H.; Madanian, S.; Malik, N.; White, D.; Russel, B.K.; Niazi, I.K. Prediction of Interstitial Glucose Levels Through Wearable Sensors Using Machine Learning. In *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*; IEEE: New York, NY, USA, 2023; pp. 1–6. [CrossRef]
15. Adams, D.; Nsugbe, E. Predictive Glucose Monitoring for People with Diabetes Using Wearable Sensors. *Eng. Proc.* **2021**, *10*, 20. [CrossRef]
16. Ali, H.; Madanain, S.; White, D.; Akhter, M.N.; Niazi, I.K. From wearable activity trackers to Interstitial Glucose: Data to Insight-A proposed scientific journey. In Proceedings of the 2024 Australasian Computer Science Week, Sydney, Australia, 29 January–1February 2024; pp. 61–64. [CrossRef]
17. Zahedani, A.D.; McLaughlin, T.; Veluvali, A.; Aghaeepour, N.; Hosseinian, A.; Agarwal, S.; Ruan, J.; Tripathi, S.; Woodward, M.; Hashemi, N.; et al. Digital health application integrating wearable data and behavioral patterns improves metabolic health. *NPJ Digit. Med.* **2023**, *6*, 216. [CrossRef] [PubMed]
18. Bent, B.; Henriquez, M.; Dunn, J.P. Cgmquantify: Python and R Software Packages for Comprehensive Analysis of Interstitial Glucose and Glycemic Variability from Continuous Glucose Monitor Data. *IEEE Open J. Eng. Med. Biol.* **2021**, *2*, 263–266. [CrossRef] [PubMed]
19. Bent, B.; Cho, P.J.; Henriquez, M.; Wittmann, A.; Thacker, C.; Feinglos, M.; Crowley, M.J.; Dunn, J.P. Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. *NPJ Digit. Med.* **2021**, *4*, 89. [CrossRef] [PubMed]
20. Qi, W.; Wang, N.; Su, H.; Aliverti, A. DCNN based human activity recognition framework with depth vision guiding. *Neurocomputing* **2021**, *486*, 261–271. [CrossRef]
21. Zhao, J.; Lv, Y.; Zeng, Q.; Wan, L. Online Policy Learning-Based Output-Feedback Optimal Control of Continuous-Time Systems. *IEEE Trans. Circuits Syst. II Express Briefs* **2022**, *71*, 652–656. [CrossRef]
22. Lehmann, V.; Föll, S.; Maritsch, M.; van Weenen, E.; Kraus, M.; Lagger, S.; Odermatt, K.; Albrecht, C.; Fleisch, E.; Zueger, T.; et al. Noninvasive Hypoglycemia Detection in People With Diabetes Using Smartwatch Data. *Diabetes Care* **2023**, *46*, 993–997. [CrossRef]
23. Huang, X.; Schmelter, F.; Seitzer, C.; Martensen, L.; Otzen, H.; Piet, A.; Witt, O.; Schröder, T.; Günther, U.; Grzegorzek, M.; et al. From Data to Insight: Predicting Interstitial Glucose in Healthy Cohort with Non-invasive Sensor Technology and Machine Learning. *arXiv* **2023**. [CrossRef]
24. Optuna—A Hyperparameter Optimization Framework. Optuna. Available online: https://optuna.org/ (accessed on 8 June 2024).
25. Liang, Y.; Elgendi, M.; Chen, Z.; Ward, R. An optimal filter for short photoplethysmogram signals. *Sci. Data* **2018**, *5*, 180076. [CrossRef] [PubMed]
26. Nabian, M.; Yin, Y.; Wormwood, J.; Quigley, K.S.; Barrett, L.F.; Ostadabbas, S. An Open-Source Feature Extraction Tool for the Analysis of Peripheral Physiological Data. *IEEE J. Transl. Eng. Heal. Med.* **2018**, *6*, 2800711. [CrossRef]
27. Lam, B.; Catt, M.; Cassidy, S.; Bacardit, J.; Darke, P.; Butterfield, S.; Alshabrawy, O.; Trenell, M.; Missier, P. Using Wearable Activity Trackers to Predict Type 2 Diabetes: Machine Learning–Based Cross-sectional Study of the UK Biobank Accelerometer Cohort. *JMIR Diabetes* **2021**, *6*, e23364. [CrossRef]
28. Interbeat Interval Filtering. Available online: https://arxiv.org/html/2406.01846v1#S3 (accessed on 22 July 2024).
29. Chandra, V.; Priyarup, A.; Sethia, D. Comparative Study of Physiological Signals from Empatica E4 Wristband for Stress Classification. In *Advances in Computing and Data Sciences*; Singh, M., Tyagi, V., Gupta, P.K., Flusser, J., Ören, T., Sonawane, V.R., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 218–229. [CrossRef]
30. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
31. Hastie, T.; Tibshirani, R.; Friedman, J. Random Forests. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Hastie, T., Tibshirani, R., Friedman, J., Eds.; Springer: New York, NY, USA, 2009; pp. 587–604. [CrossRef]
32. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
33. Hastie, T.; Tibshirani, R.; Friedman, J. Boosting and Additive Trees. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Hastie, T., Tibshirani, R., Friedman, J., Eds.; Springer: New York, NY, USA, 2009; pp. 337–387. [CrossRef]
34. Zhang, H. The Optimality of Naive Bayes. AAAI. Available online: https://aaai.org/papers/flairs-2004-097/ (accessed on 5 June 2024).
35. Wang, H.; Hu, D. Comparison of SVM and LS-SVM for regression. In *2005 International Conference on Neural Networks and Brain*; IEEE: New York, NY, USA, 2005; pp. 279–283.
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **1994**, *16*, 235–240. [CrossRef]

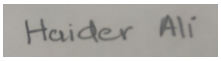
Contributions

As Co-author, I hereby approve and declare that my role in this study and percentage of contribution, as indicated below, is representative of my actual contribution and I hereby give my consent that this work may be published as part of this PhD thesis.

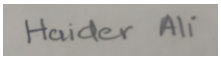
Chapter 2: Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning

Author	Contribution	Percentage	Signature
H.A.	Concept and design of the study Data Analysis and interpretation Writing	80%	
S.M.	Concept and design of the study Reviewing and editing	7%	
D.W.	Concept and design of the study Reviewing and editing	7%	
I.K.N.	Reviewing and editing	6%	

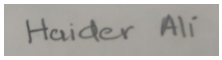
Chapter 3: Digital Biomarkers from smart watches and food logs for Interstitial Glucose Prediction: A Systematic Review

Author	Contribution	Percentage	Signature
H.A.	Concept and design of the study Data Analysis and interpretation Writing	80%	
S.M.	Concept and design of the study Reviewing and editing	7%	
D.W.	Concept and design of the study Reviewing and editing	7%	
I.K.N	Reviewing and editing	6%	

Chapter 4: Comparison of Machine Learning Models for Predicting Interstitial Glucose Using Smart Watch and Food Log

Author.	Contribution	Percentage	Signature
H.A.	Concept and design of the study Data collection Data Analysis and interpretation Writing	80%	
S.M.	Concept and design of the study Reviewing and editing	6%	
D.W.	Concept and design of the study Reviewing and editing	7%	
I.K.N	Data collection	7%	

Chapter 5: Sleep Features for Prediction of Interstitial Glucose

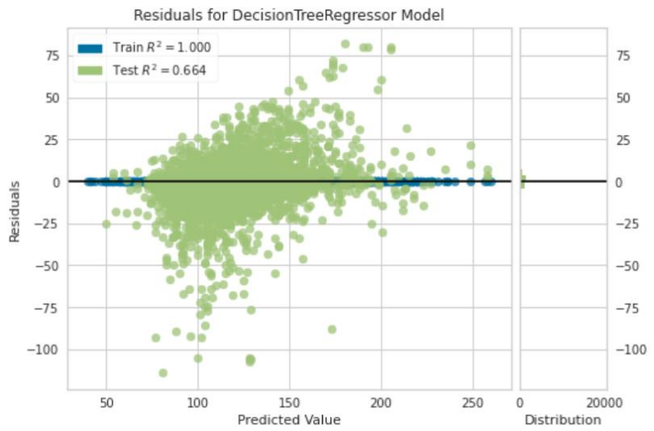
Author	Contribution	Percentage	Signature
H.A.	Concept and design of the study Data collection Data Analysis and interpretation Writing	80%	
S.M.	Concept and design of the study Reviewing and editing	7%	
D.W	Reviewing and editing	6%	
I.K.N	Concept and design of the study Reviewing and editing	6%	

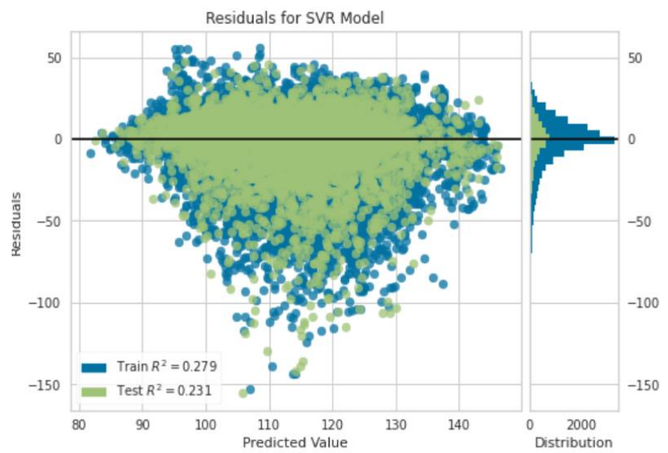
Supplementary Materials

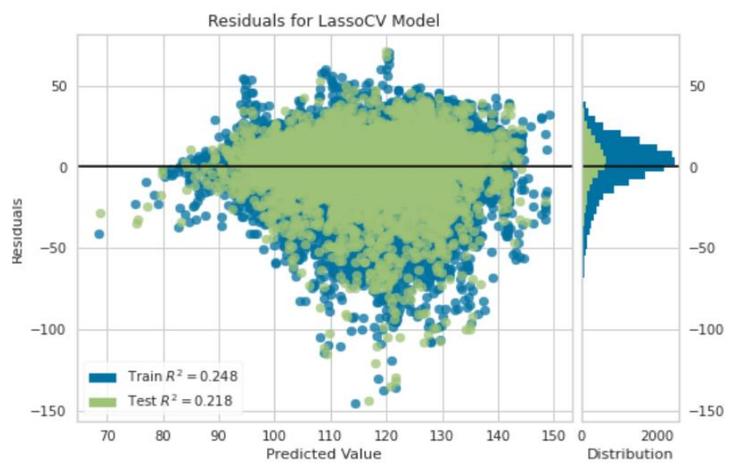
Table S1: Demographics of the participants in the dataset.

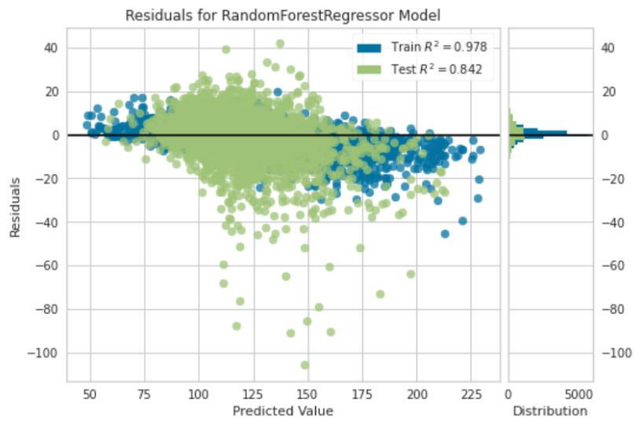
ID	Gender	HbA1c
13	MALE	5.7
1	FEMALE	5.5
3	FEMALE	5.9
4	FEMALE	6.4
5	FEMALE	5.7
2	MALE	5.6
6	FEMALE	5.8
7	FEMALE	5.3
8	FEMALE	5.6
10	FEMALE	6.0
9	MALE	6.1
11	MALE	6.0
12	MALE	5.6
14	MALE	5.5
15	FEMALE	5.5
16	MALE	5.5

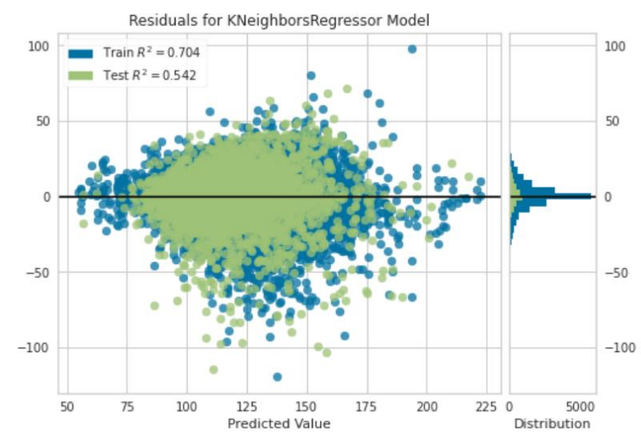
Figure S1: Performance comparison of the regression models in Chapter 4.

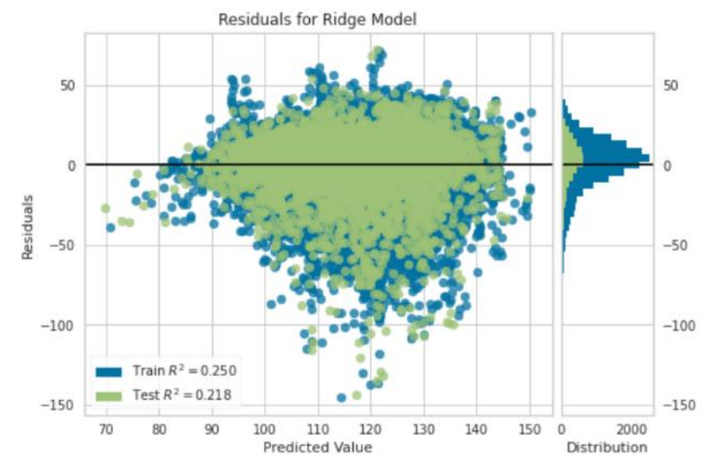


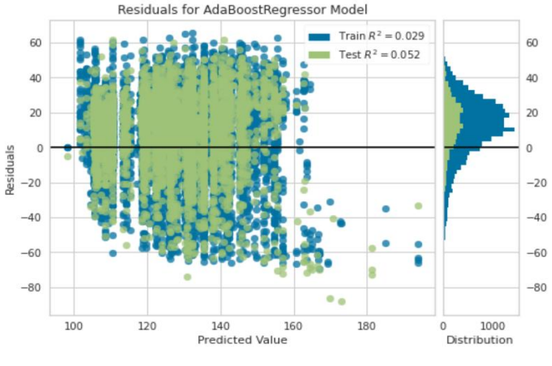












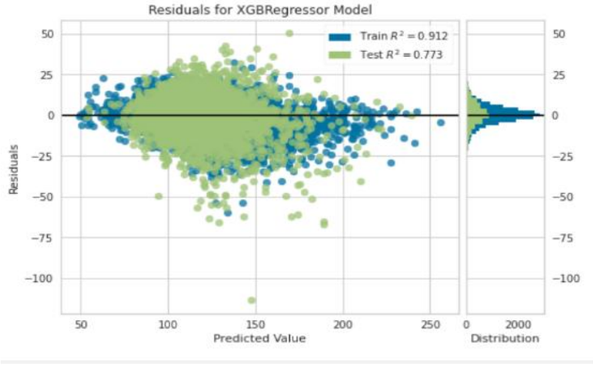
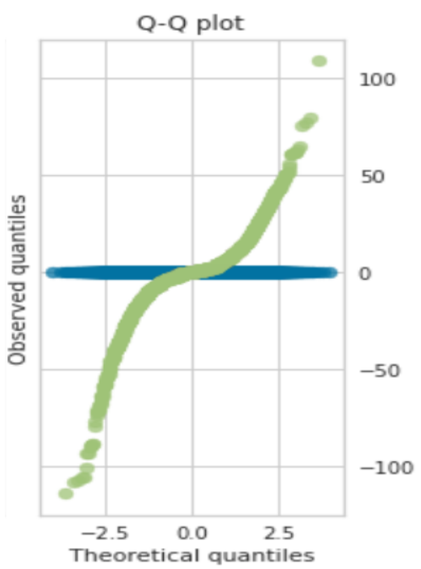
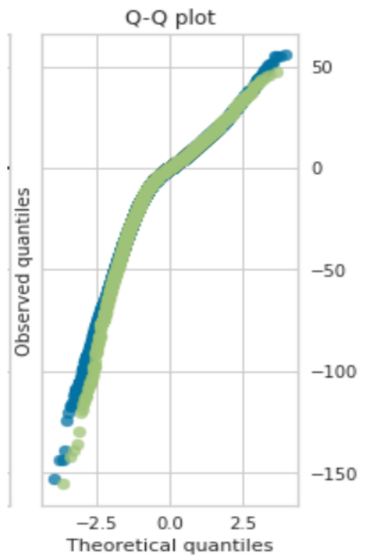


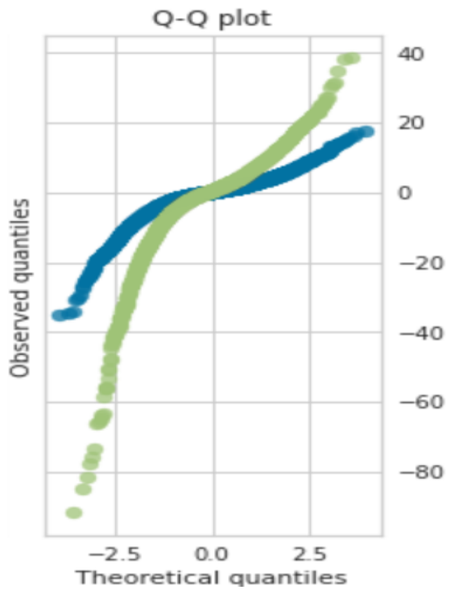
Figure S2: QQ plot of the Regression models in Chapter 4



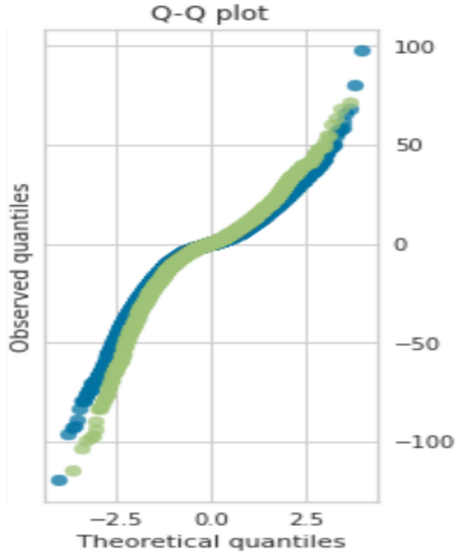
DT



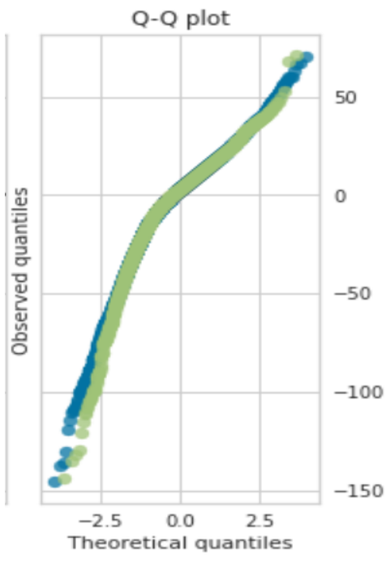
SVR



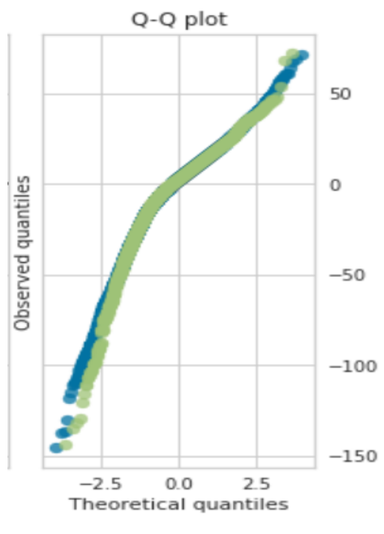
RF



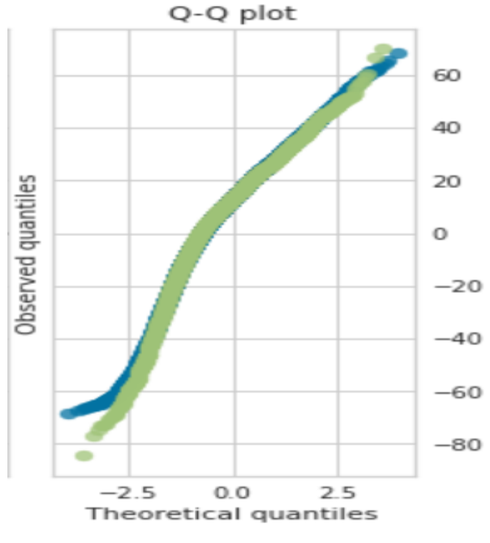
KNN



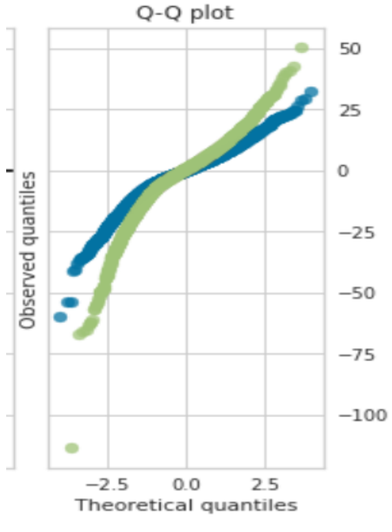
LASSOCV



RIDGE

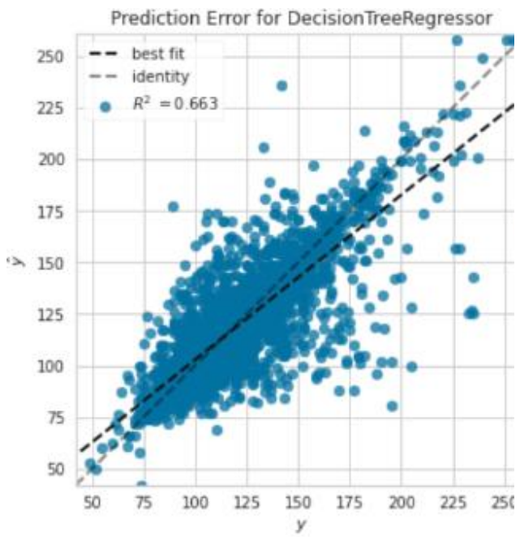


ADABOOST

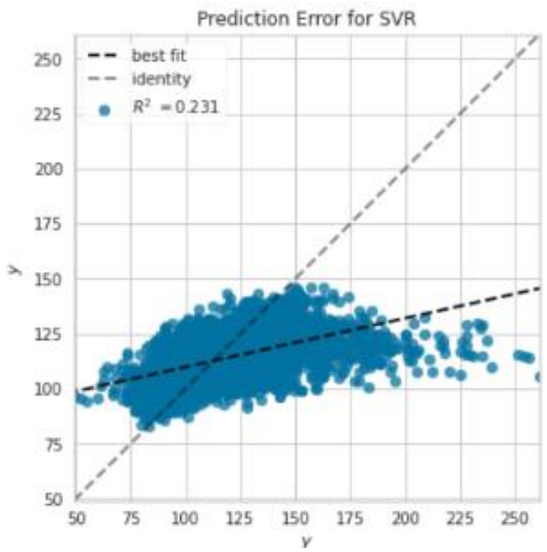


XGBOOST

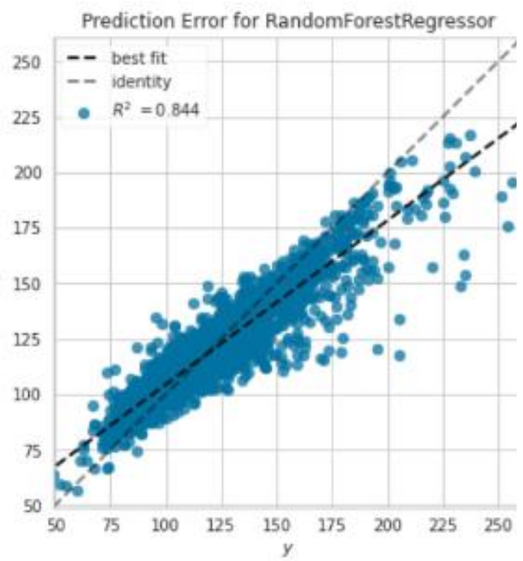
Figure S3: Prediction Error of the Regression Models for Chapter 4



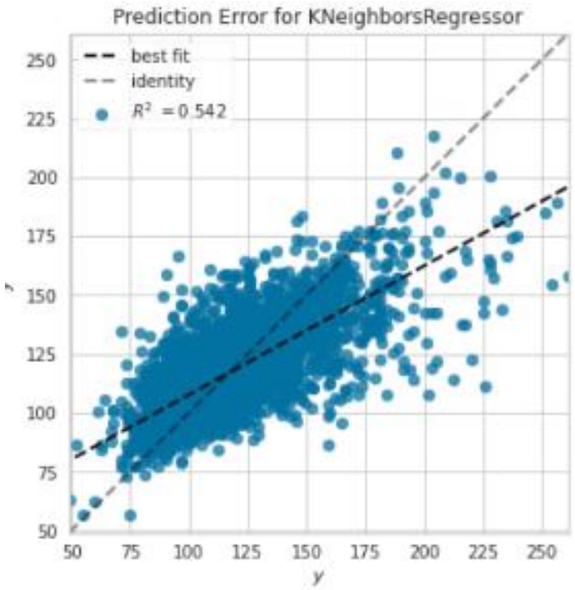
DT



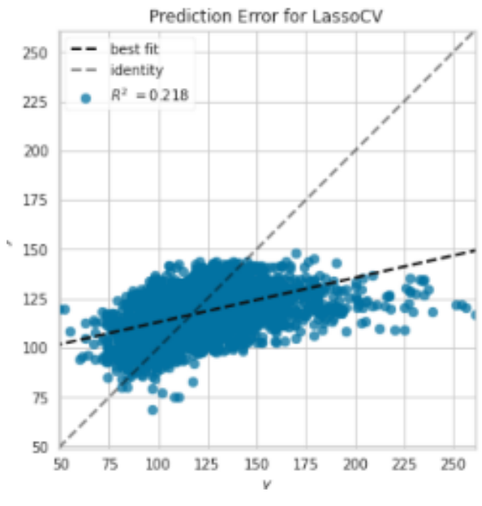
SVR



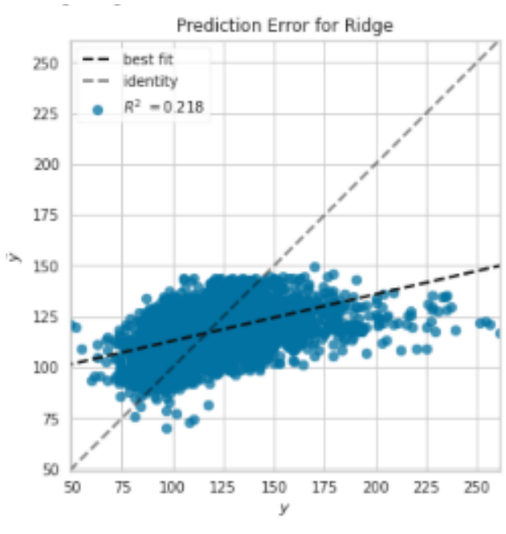
RF



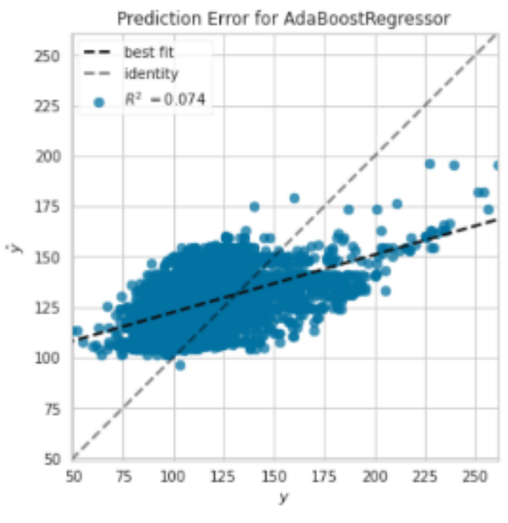
KNN



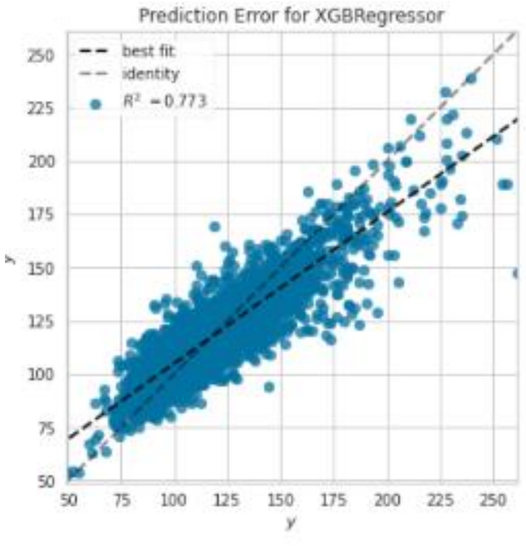
LASSOCV



RIDGE



ADABOOST



XGBOOST

Figure S4: Regression Model performance (kfold k=10) for Chapter 4.

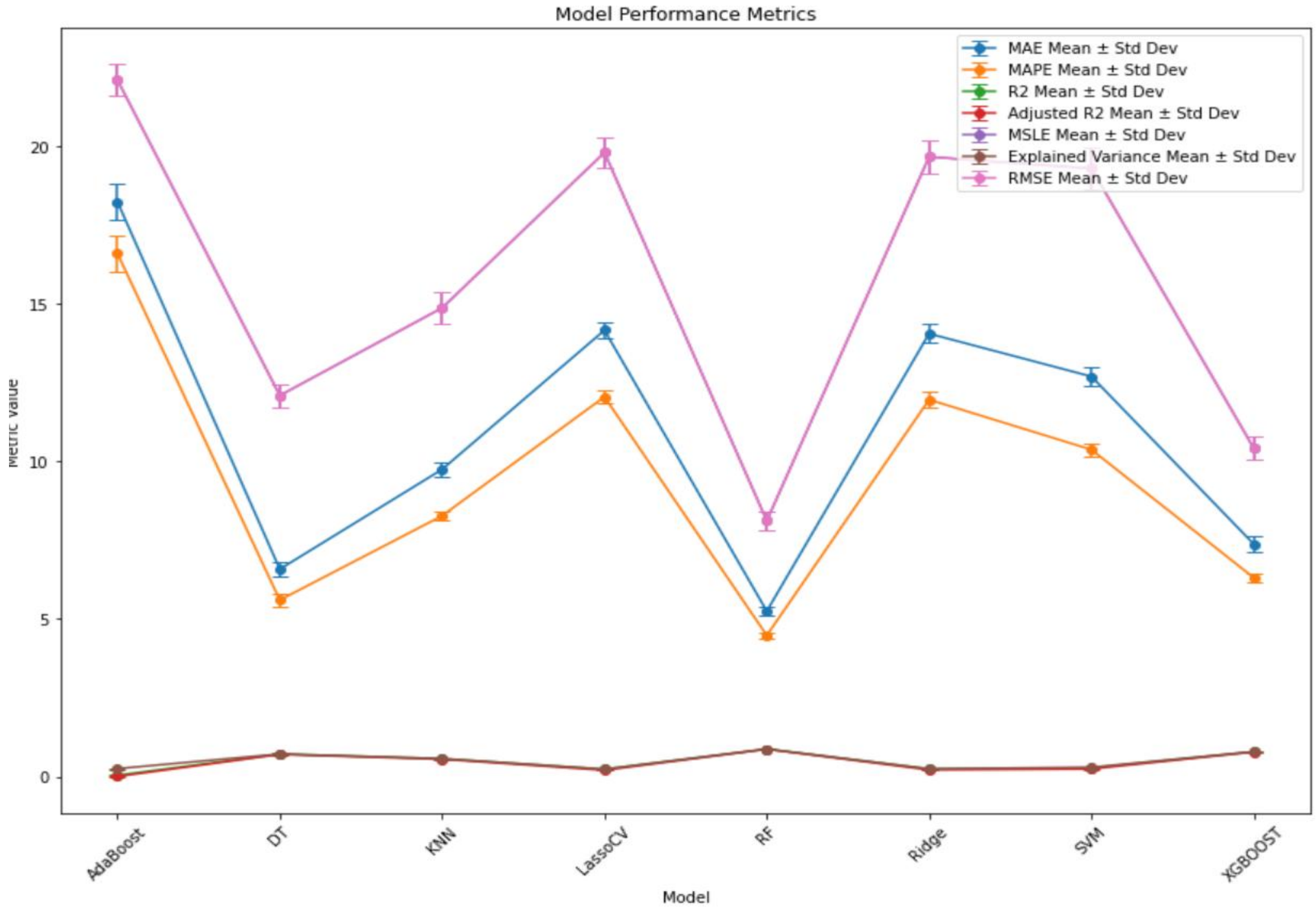


Figure S5: t-SNE plot of the features, here purple represents low glucose, green represents high glucose and yellow represents low glucose

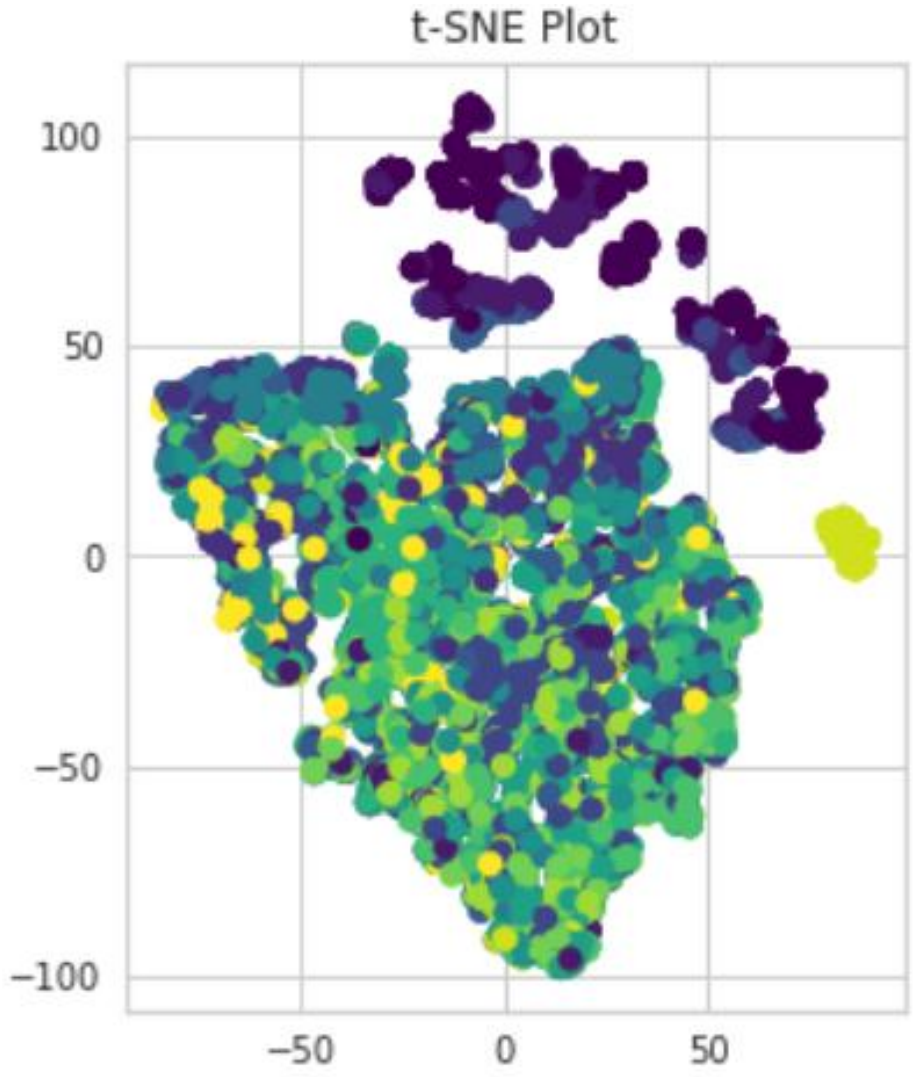
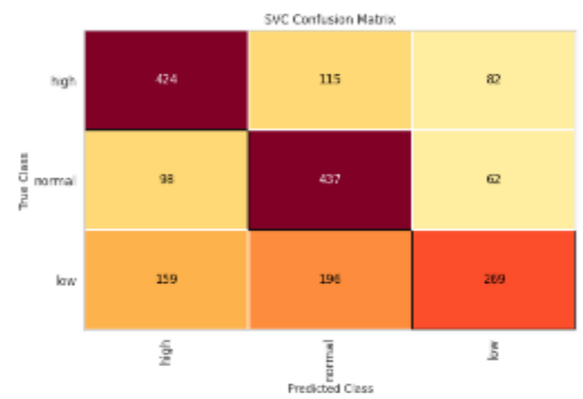
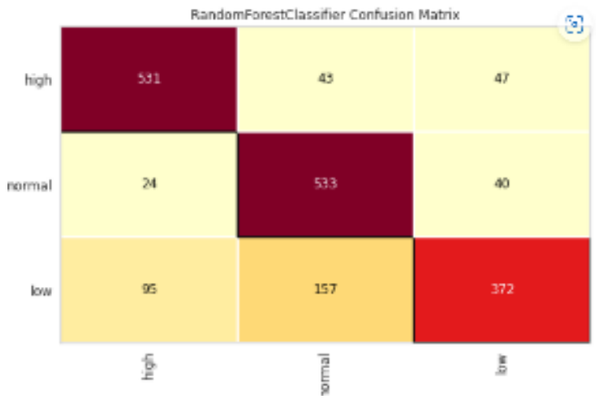


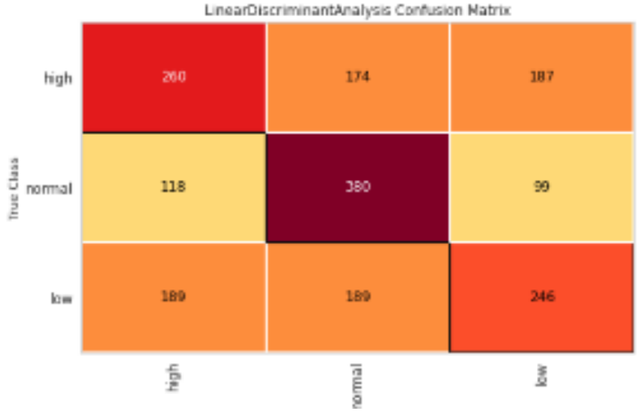
Figure S6: Confusion Matrix of the Classification Models for Chapter 4



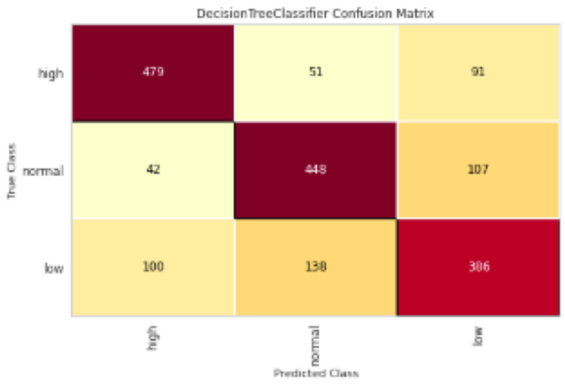
SVM



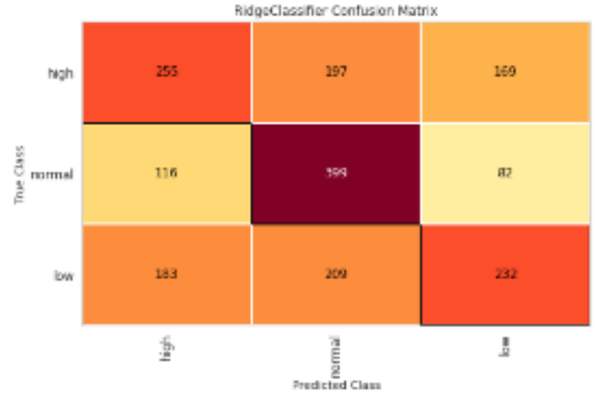
RF



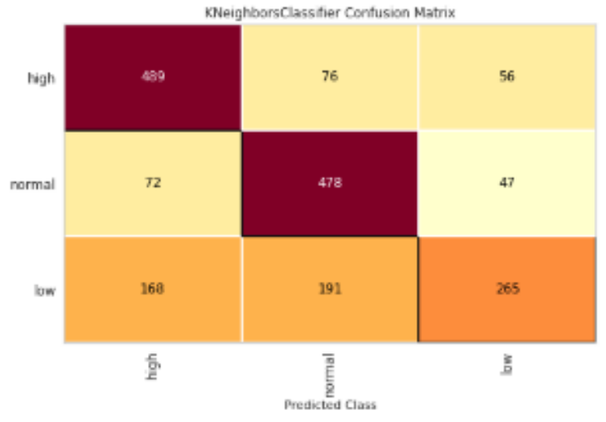
LDA



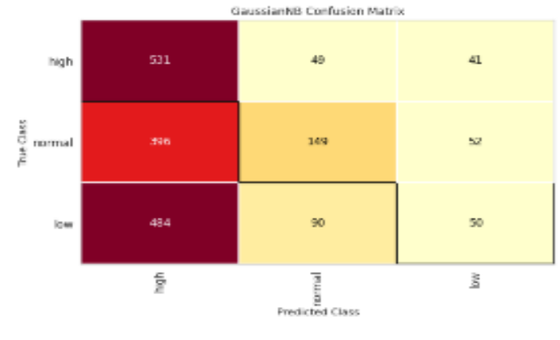
DT



RIDGE

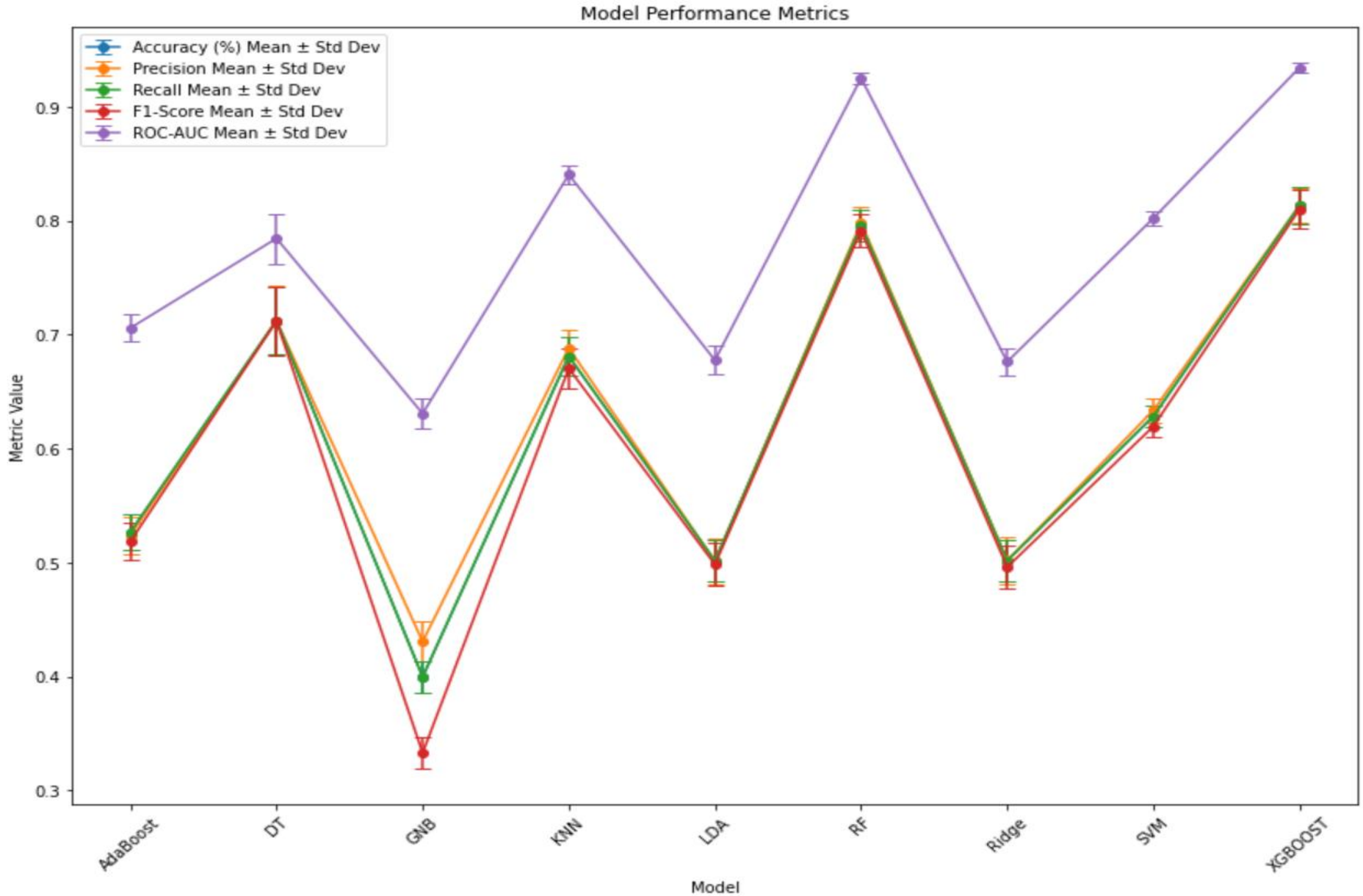


KNN



GNB

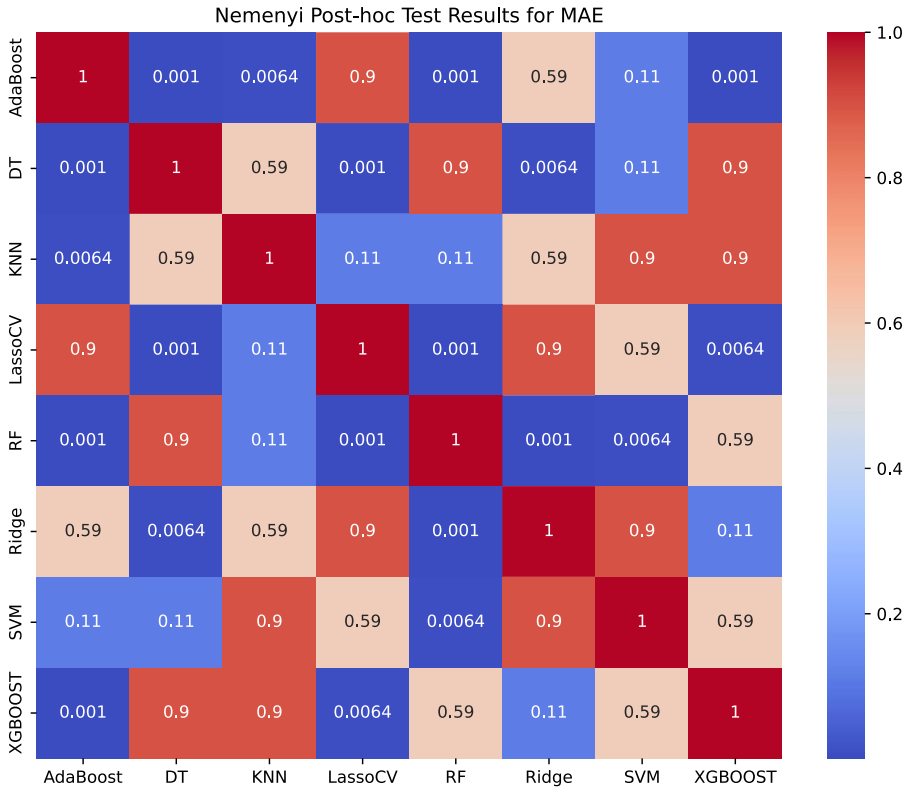
Figure S7: Classification Model performance (kfold k=10).



Comparison of significance of performance parameters for Regression Chapter 4

Firstly, a Friedman Statistic is calculated and then a Nemenyi post hoc analysis is performed on the results of Friedman test**.**

Figure S8: Nemenyi Post-hoc results for MAE



Explanation:

The Nemenyi post-hoc test results for Mean Absolute Error (MAE) reveal significant and non-significant differences in model performance. Significant differences ($p < 0.05$) were observed between AdaBoost and Decision Tree (p=0.001), AdaBoost and KNN (p=0.006), AdaBoost and Random Forest (p=0.001), AdaBoost and XGBoost (p=0.001), Decision Tree and LassoCV (p=0.001), Decision Tree and Ridge (p=0.006), Random Forest and LassoCV (p=0.001), Random Forest and Ridge (p=0.001), SVM and Random Forest (p=0.006), and XGBoost and LassoCV (p=0.006). No significant differences ($p \geq 0.05$) were found between AdaBoost and LassoCV, AdaBoost and Ridge, AdaBoost and SVM, Decision Tree and KNN, Decision Tree and Random Forest, Decision Tree and SVM, Decision Tree and XGBoost, KNN and LassoCV, KNN and Random Forest, KNN and Ridge, KNN and SVM, KNN and XGBoost, LassoCV and Ridge, LassoCV and SVM, Random Forest and XGBoost, Ridge and SVM, Ridge and XGBoost, and SVM and XGBoost.

Figure S9: Nemenyi Post-hoc results for Adjusted R2

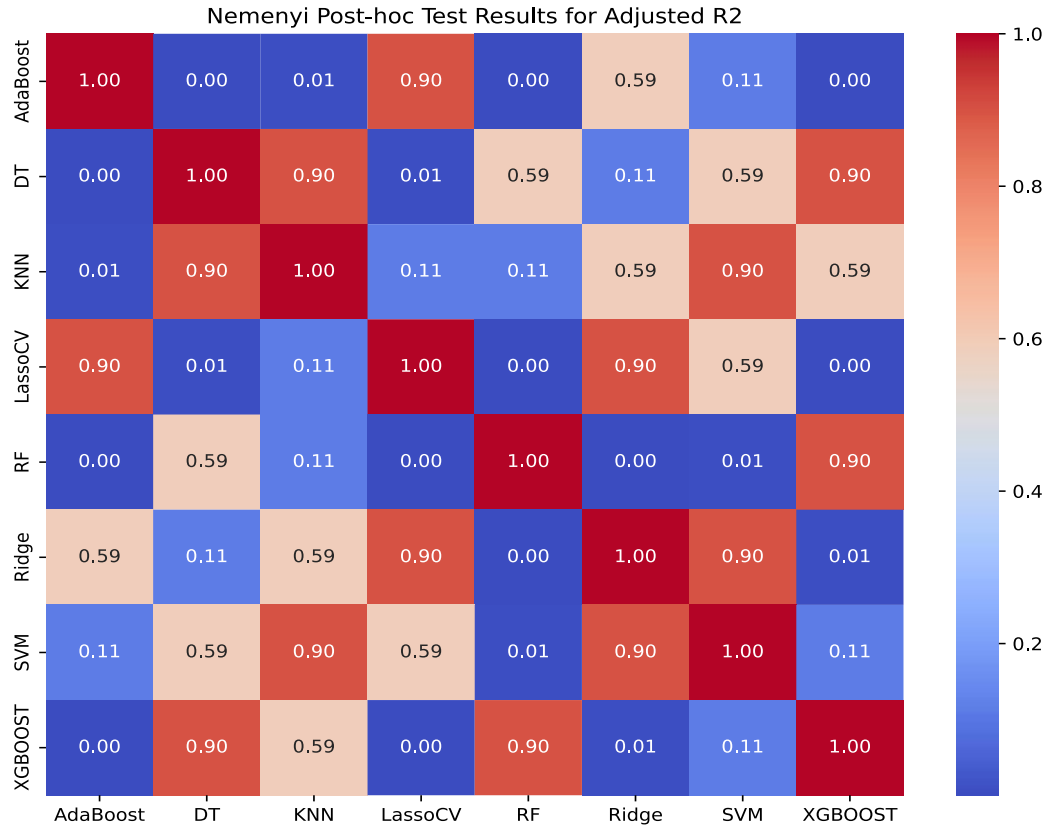


Figure S10: Nemenyi Post-hoc results for Explained Variance

Figure S11: Nemenyi Post-hoc results for MSLE

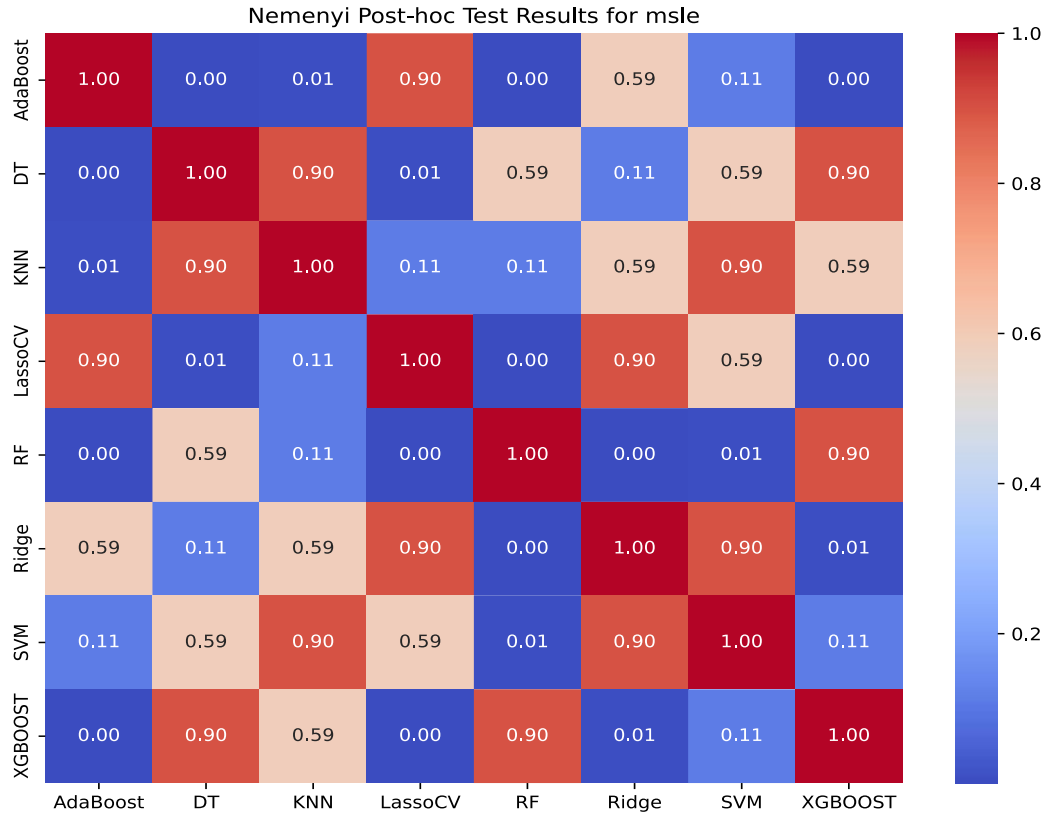


Figure S12: Nemenyi Post-hoc results for MAE

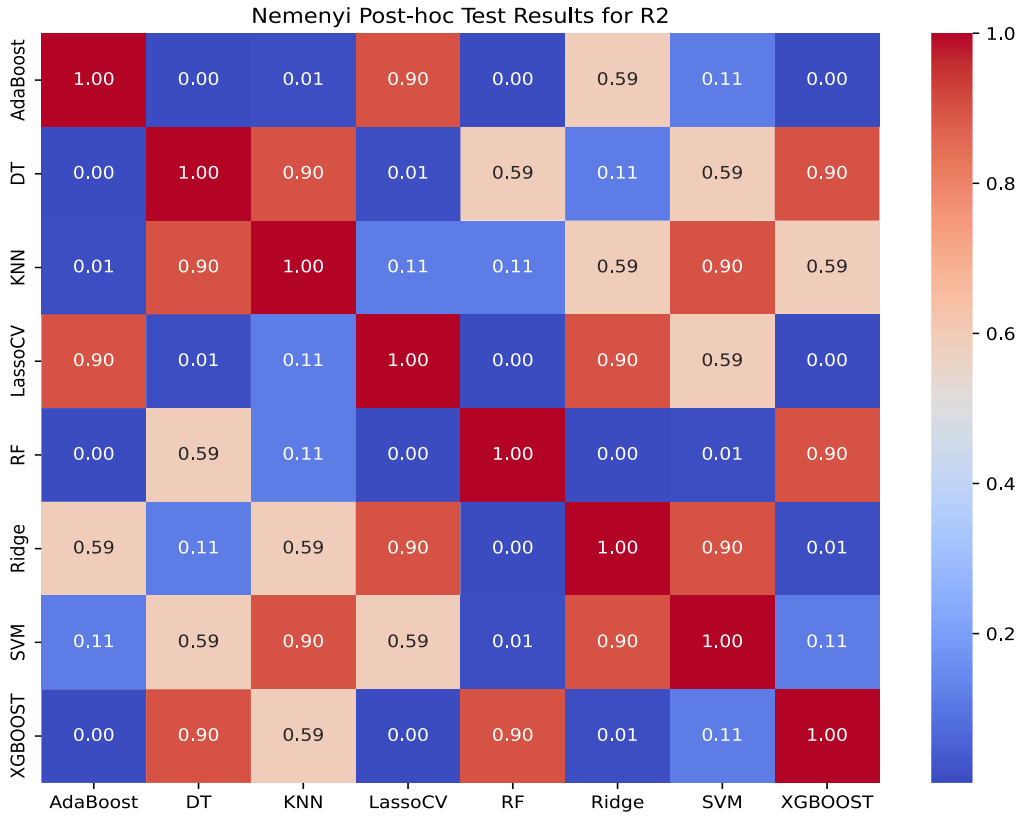


Figure S13: Nemenyi Post-hoc results for Accuracy

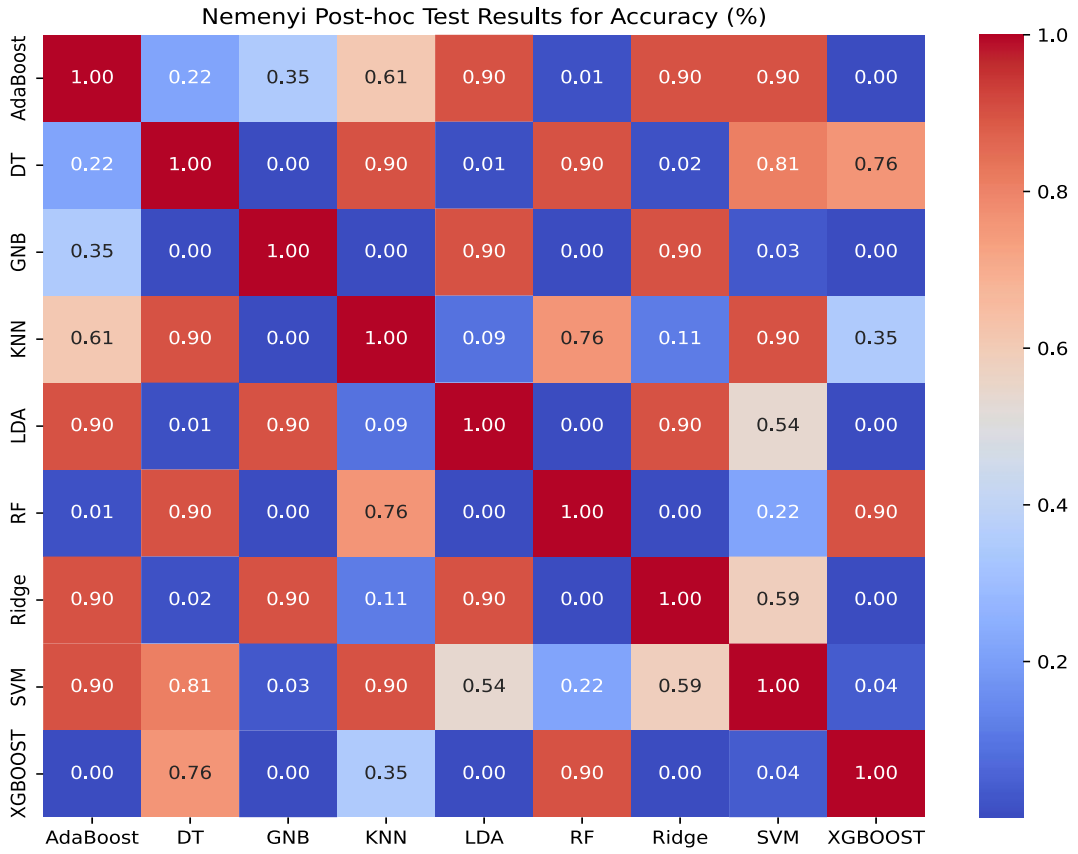


Figure S14: Nemenyi Post-hoc results for precision

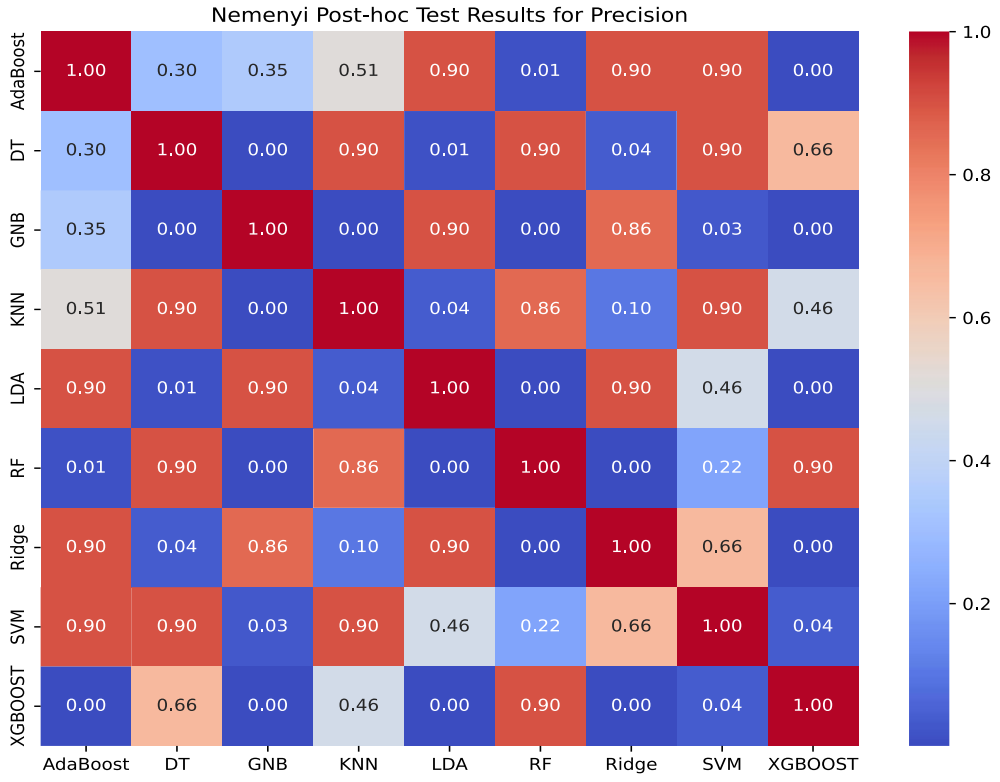


Figure S15: Nemenyi Post-hoc results for F-1 Score

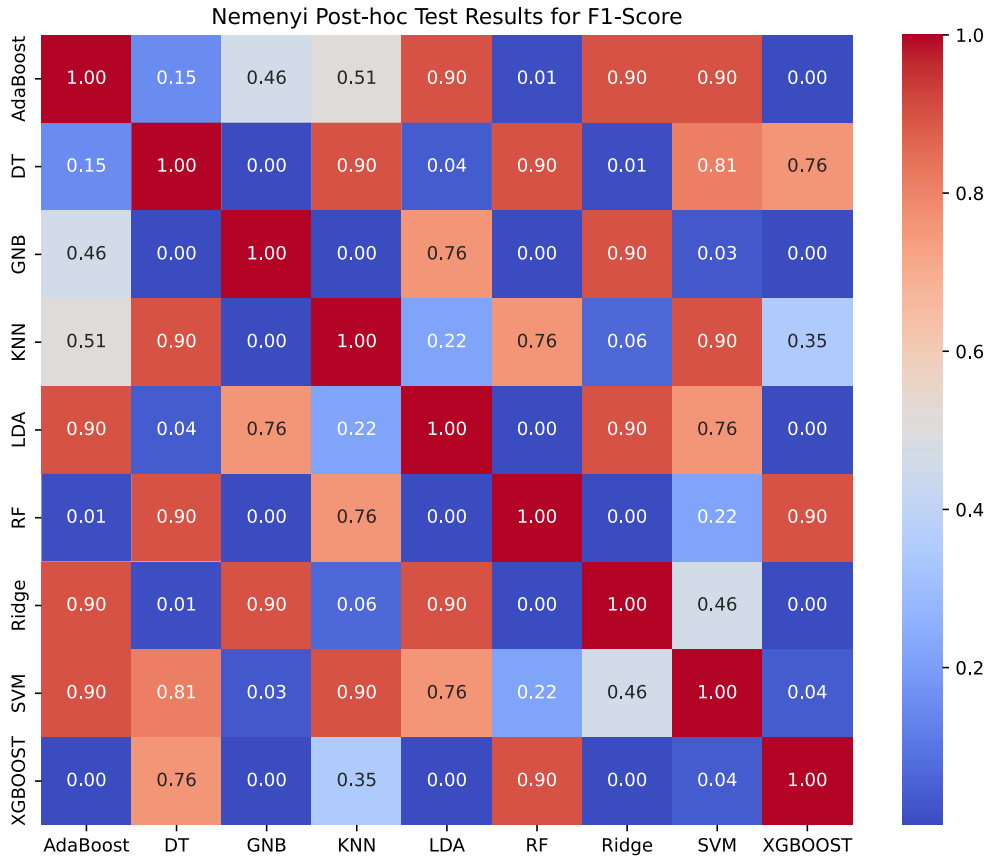


Figure S16: Nemenyi Post-hoc results for Recall

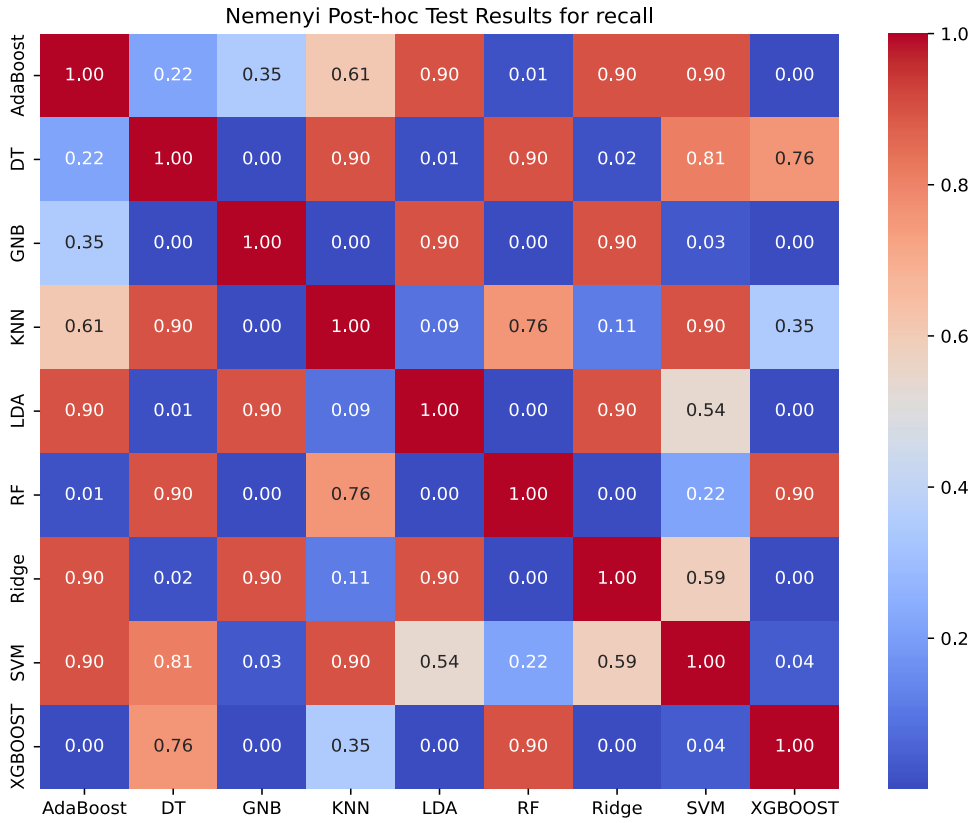


Figure S17: Nemenyi Post-hoc results for ROC-AUC